Model Based System Analysis Techniques to Determine Propagation Paths of Functional Insufficiencies in Software Intensive Systems

Thesis approved by the

Department of Computer Science
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

for the award of the Doctoral Degree

Doctor of Engineering (Dr.-Ing.)

to

M.Sc. Ahmad ADEE

Date of Defense:	29.09.2023
Dean:	Prof. Dr. Christoph GARTH
Reviewer:	Prof. Dr. Peter LIGGESMEYER
Reviewer:	Prof. Dr. Mario TRAPP
Reviewer:	Dr. Andreas HEYL

DE-386

"If I have seen further, it is by standing on the shoulders of giants."

Isaac Newton

Abstract

Highly Automated Driving (HAD) vehicles represent complex and safety critical systems. They are deployed in an open context i.e., an intricate environment which undergoes continual changes. The complexity of these systems and insufficiencies in sensing and understanding the open context may result in unsafe and uncertain behaviour. The safety critical nature of the HAD vehicles requires modelling of root causes for unsafe behaviour and their mitigation to argue sufficient reduction of residual risk.

Standardization activities such as ISO 21448 provide guidelines on the Safety Of The Intended Functionality (SOTIF) and focus on the analysis of performance limitations under the influence of triggering conditions that can lead to hazardous behaviour. SOTIF references traditional safety analyses methods e.g., Failure Mode and Effect Analysis (FMEA) and Fault Tree Analysis (FTA) to perform safety analysis. These analyses methods are based on certain assumptions e.g., single point failure in FMEA and independence of basic events in FTA. Moreover, these analyses are generally based on expert knowledge i.e., data-based models or hybrid approaches (expert and data) are seldom practised. The resulting safety model is fixed i.e., it is generally seen as a one-time artefact. Open context environment may contain triggering conditions which may not be evident to the expert. Open context also evolves over time and new phenomena may emerge.

This thesis explores the applicability of the traditional safety analyses techniques to provide safety models for HAD vehicles operating in the open context, under the light of modelling assumptions taken by traditional safety analyses techniques. Moreover, incorporating uncertainties into safety analyses models is also explored. An explicit distinction between the inherent uncertainty of a probabilistic event (aleatory) and uncertainty due to lack of knowledge (epistemic) is made to formalize models to perform SOTIF analysis. A further distinction is made for conditions of complete ignorance and termed as ontological uncertainty. The distinction is important as for HAD vehicles operating in open context the ontological uncertainty can never be completely disregarded.

This thesis proposes a novel framework of SOTIF to model, estimate and discover triggering conditions relevant to performance limitations. The framework provides the ability to model uncertainties while also providing a hybrid approach i.e., supporting inclusion of expert knowledge as well as data driven engineering processes. Two representative algorithms are provided to support the framework. Bayesian Network (BN) and p-value hypothesis testing are utilised in this regard. The framework is implemented on a real-world case study in which LIDARs based perception systems are used as vehicle detection system.

Acknowledgements

I would like to communicate my gratitude to my advisor, *Prof. Dr. Peter Liggesmeyer* for his endless support, invaluable guidance, and perseverance throughout this research journey. His mentorship has been instrumental in shaping this thesis and my academic growth.

I am immensely thankful to the committee members and reviewers, *Prof. Dr. Peter Liggesmeyer*, *Prof. Dr. Mario Trapp*, *Dr. Andreas Heyl* and *Prof. Dr. Klaus Schneider* for their insightful feedback and constructive criticism, which greatly enriched the quality of this work.

I would also like to thank my PhD supervisors at Robert Bosch GmbH, *Dr. Roman Gansch* and *Dr. Peter Munk*. Your guidance and belief in me have been invaluable.

I extend my heartfelt appreciation to *Robert Bosch GmbH* for providing the resources and conducive research environment essential for this study.

I am beholden to my colleagues and fellow researchers, *Markus Schweizer*, *Dr. Andreas Rohatschek*, *Dr. Arne Nordmann*, *Dr. Lydia Gauerhof*, *Arut Prakash Kaleeswaran* and *Dr. Eike Martin Thaden* whose stimulating discussions and encouragement kept me motivated. Your camaraderie has made this academic endeavor more enriching.

Thanks to all my colleagues and professors from the *Safer and Autonomous Systems (SAS)* project for the great exchange and amusing moments that we spent together.

Special thanks to my family and friends for their unwavering support, understanding, and encouragement. Your belief in me sustained my determination during challenging times.

Finally, I would like to mention European Union who funded this thesis and the overall project under the Marie Skłodowska-Curie grant agreement No 812.788 (MSCA-ETN SAS). This dissertation reflects only the authors' view, exempting the European Union from any liability. Project website: http://etn-sas.eu/.

This research would not have been possible without the collective support and encouragement of all those mentioned above. Thank you from the depths of my heart.

Ahmad ADEE
29th September 2023

Contents

List of Figures

List of Tables

List of Abbreviations

AEB	**A**utomated **E**mergency **B**raking
ALARP	**A**s **L**ow **A**s **R**easonably **P**ossible
BBA	**B**asic **B**elief **A**ssignment
BN	**B**ayesian **N**etwork
BSI	**B**ritish **S**tandard **I**nstitute
CBL	**C**onditional **B**elief **L**ikelihood
CBN	**C**ausal **B**ayesian **N**etwork
CBT	**C**onditional **B**elief **T**able
CCET	**C**onfounding **C**ausal **E**dge **T**rail
CPLM	**C**onditional **P**erformance **L**imitation **M**ap
CPS	**C**yber **P**hysical **S**ystem
CTA	**C**ause **T**ree **A**nalysis
CS	**C**ausal **S**cenarios
DAG	**D**irected **A**cyclic **G**raph
DCET	**D**irect **C**ausal **E**dge **T**rail
DNN	**D**eep **N**eural **N**etwork
DST	**D**empster (and) **S**hafer **T**heory
E/E	**E**lectrical and **E**lectronic
EEN	**E**xtended **E**vidential **N**etwork
EN	**E**vidential **N**etwork
EMI	**E**lectro-**M**echanical **I**nterference
ESD	**E**vent **S**equence **D**iagram
ET	**E**vidence **T**heory
ETA	**E**vent **T**ree **A**nalysis
FP	**F**alse **P**ositive
FN	**F**alse **N**egative
FFTA	**F**uzzy **F**ault **T**ree **A**nalysis
FTA	**F**ault **T**ree **A**nalysis
FMEA	**F**ailure **M**ode and **E**ffect **A**nalysis
GAMAB	**G**lobalement **A**u **M**oins **A**ussi **B**on
HARA	**H**azard **A**nalysis and **R**isk **A**ssessment
HAD	**H**ighly **A**utomated **D**riving
HMI	**H**uman **M**achine **I**nterface
HAZOP	**Haz**ard and **Op**erability study
ISO	**I**nternational **O**rganization for **S**tandardization
KL	**K**ullback-**L**eibler
LIDAR	**L**ight **D**etection **A**nd **R**anging
MLE	**M**aximum **L**ikelihood **E**stimator
MSE	**M**ean **S**quared **E**rror
NTC	**N**ovel **T**riggering **C**ondition
ODD	**O**perational **D**esign **D**omain
OEM	**O**riginal **E**quipment **M**anufacturer

PAS	**P**ublicly **A**ccessible **S**pecification
PGM	**P**robabilistic **G**raphical **M**odel
PLM	**P**erformance **L**imitation **M**ap
PPMI	**P**ositive **P**ointwise **M**utual **I**nformation
RADAR	**Ra**dio **D**etection **A**nd **R**anging
RPN	**R**isk **P**riority **N**umber
RSS	**R**elevant **S**cene **S**core
SAE	**S**ociety of **A**utomotive **E**ngineers
SOTIF	**S**afety **O**f **T**he **I**ntended **F**unctionality
STAMP	**S**ystems-**T**heoretic **A**ccident **M**ode and **P**rocesses
STPA	**S**ystem **T**heoretic **P**rocess **A**nalysis
TP	**T**rue **P**ositive
TTC	**T**ime **T**o **C**ollusion
TTM	**T**ime **T**o **M**aterialization
UCA	**U**nsafe **C**ontrol **A**ction
UL	**U**nderwriter **L**aboratories
V&V	**V**erification & **V**alidation
V2V	**V**ehicle to **V**ehicle
V2X	**V**ehicle to Everything

Publications

This thesis is based on and in parts taken from our following publications.

Gansch, Roman, and Ahmad Adee. "System theoretic view on uncertainties." In 2020 *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1345-1350. IEEE, 2020. [41]

Adee, A., P. Munk, R. Gansch, and P. Liggesmeyer. "Uncertainty Representation with Extended Evidential Networks for Modeling Safety of the Intended Functionality (SOTIF)." In *European Safety and Reliability Conference (ESREL2020)*, pp. 4148-4156. 2020. [3]

Adee, Ahmad, Roman Gansch, and Peter Liggesmeyer. "Systematic modeling approach for environmental perception limitations in automated driving." In 2021 *17th European Dependable Computing Conference (EDCC)*, pp. 103-110. IEEE, 2021. [1]

Adee, Ahmad, Roman Gansch, Peter Liggesmeyer, Claudius Glaeser, and Florian Drews. "Discovery of Perception Performance Limiting Triggering Conditions in Automated Driving." In 2021 *5th International Conference on System Reliability and Safety (ICSRS)*, pp. 248-257. IEEE, 2021. [2]

Intellectual Property Rights

The following intellectual property rights were filed during the time period of this dissertation.

Model Based Safety Analysis (MBSA) for Safety Of The Intended Functionality (SOTIF).

Mechanism to derive and configure a self-adaptive system from its uncertainty analysis pertaining to SOTIF.

Implementation of SOTIF related capabilities for perception of autonomous systems through parameter learning.

Calibratable SOTIF interface for sensor systems.

Calibration of onboard sensors of AD systems for SOTIF by external and infrastructure systems.

Configuration of onboard sensors of AD systems for SOTIF by external and infrastructure systems.

Identification of novel causal factors (triggering conditions) using probabilistic graphical models and hypothesis testing.

Miscellaneous Contributions

The following miscellaneous contributions were made during the research of this dissertation.

Blog: Uncertainties and Automated Driving.

Blog: The Case for Safety Analysis.

EU Report: Completeness of Model-based System Analysis when Dealing with Functional Insufficiencies

For/Dedicated to/To my…

1

Introduction

"Every beginning is difficult, holds in all sciences."

– Karl Marx, *German Philosopher*

This chapter positions the dissertation in the safety engineering domain and provides argumentation on the relevancy of the research conducted. First, the trends in the automotive industry are introduced and the impacts of those trends on achieving safety are discussed (Sec. 1.1). Definitions that are relevant to understand this section are provided. Thesis goals are summarized by formulating three main research questions (Sec. 1.2). In Sec.1.3, the scientific contributions are summarized in relation to the formulated research questions. The scientific contributions are followed by the validation of the proposed approach (Sec. 1.4). Lastly, the outline of the thesis is given (Sec. 1.5).

1.1 Motivation and Problem Statement

Highly Automated Driving (HAD)[1] vehicles are envisioned to revolutionize road mobility and are foreseen to bring positive social, economic and environmental impacts. Society of Automotive Engineers (SAE) defines six levels of driving automation ranging from simpler automation function to fully autonomous driving [97]. Whether it is the simple automation function of cruise control or the vision of fully autonomous driving, the advancement in road mobility aims to produce intelligent, fuel efficient and safer road vehicles of the future [81, 25]. While lower level of automation still requires human driver intervention, the higher level of automation may completely take the human driver out of the loop [97]. Numerous Original Equipment Manufacturers (OEMs) such as Mercedes-Benz, General Motors and Audi as well as technology companies such as Apple, Qualcomm and Amazon have invested in the development of HAD vehicles in some capacity which confirms the impact of such paradigm in the industry.

With all the benefits that HAD vehicles and related automated driving functions promise to bring, they also bring in enormous potential risk [16, 62]. HAD vehicles are *safety critical systems*. Safety critical systems are those systems that have the potential to injure or kill a human, damage the property or cause environmental harm [115]. For example, failure of Electrical/Electronic (E/E) equipment in automotive vehicles may lead to harm to humans or the environment. International Organization for Standardization (ISO) addresses these types of failures under the term *functional safety* and has addressed it in the functional safety standard, ISO 26262 [55]. The standard defines functional safety as follows.

[1] In this thesis, the term Highly Automated Driving (HAD) is deliberately used to cover vast amount of automated driving functions irrespective to their SAE level [97].

Definition 1 *The absence of unreasonable risk due to hazards caused by malfunctioning behaviour of the E/E system is termed as functional safety [55].*

HAD vehicle rely on sensing the external environment to build *situational awareness*[2]. The *intended functionality*[3] and its implementation for such a system may also cause hazardous behaviour, despite the fulfillment of functional safety addressed in ISO 26262. The causes may include but are not limited to [54].

1. Perceiving the environment incorrectly
2. Lack of robustness of the functions and system against adverse environmental conditions
3. Unexpected behaviour emanating from decision algorithm

In order to address these causes, the ISO published ISO 21448 Road vehicles Safety of the Intended Functionality (SOTIF). ISO 21448 defines SOTIF as follows.

Definition 2 *The absence of the unreasonable risk due to hazards resulting from functional insufficiency of the intended functionality is termed as SOTIF [54].*

The standard defines the *functional insufficiency* as follows.

Definition 3 *Insufficiency of specifications*[a] *and performance limitations*[b] *is termed as functional insufficiency [54].*

[a] Specification, possibly incomplete, leading to hazardous behaviour in combination with one or more triggering conditions [54].
[b] Limitation of the technical capability leading to hazardous behaviour in combination with one or more triggering conditions [54].

SOTIF intends to provide guidance applicable on the design, Verification and Validation (V&V) measures needed to achieve SOTIF for lower level of automated driving functions (up to SAE level 2 [97]). Specifically, SOTIF is applied to the intended functionality where proper situational awareness is critical to safety and the situational awareness is derived from complex sensors and processing algorithms [54]. While the standard can be considered for higher level of automation (SAE level 3 and above [97]), additional measures might be required.

Unlike functional failures that are addressed by ISO 26262 [55], SOTIF argues that *triggering conditions*[4] (causal factors e.g., environmental conditions, road conditions) and their combinations thereof may activate functional insufficiencies which can result in hazardous behaviour at vehicle level [54].

Hazards emanating from lack of functional safety and SOTIF are generally managed with the methods and tools of safety engineering. In the automotive industry, safety analyses methods include but are not limited to Hazard and Operability Analysis (HAZOP), Event Tree Analysis (ETA), Failure Mode and Effect Analysis (FMEA) and Fault Tree Analysis (FTA). The proven in use argumentation is reflected in some automotive safety standards e.g., ISO 26262 [55]. SOTIF utilizes FMEA and FTA to perform safety analyses on the HAD vehicle functions. It also references the use of System Theoretic Process Analysis (STPA) [54].

[2] Understanding of the situation [54]
[3] Specified functionality [54].
[4] "Specific conditions of a scenario that serve as an initiator for a subsequent system reaction leading to hazardous behaviour [54]."

HAD vehicles are complex systems operating in *open context* [19]. The open context is the unstructured, public real-world environments in which the HAD vehicles are deployed. The complexities of these systems and the open context nature of the environment they are deployed in result in unsafe and uncertain behaviour due to functional insufficiencies in sensing and understanding the operational environment [19]. Multiple triggering conditions may exist and have an unknown causal impact on the functional insufficiencies i.e., the resulting causal model that can be defined between triggering conditions and functional insufficiencies may not be deterministic [41, 3, 1]. Moreover, the chosen causal model may result in *confounding* phenomena, resulting in a spurious relation between a dependent and independent variable [2]. Due to the open context and general lack of knowledge about the context and system, modelling all the triggering conditions becomes a particularly challenging task. Thus, the identification of novel triggering conditions appears to be specifically important and challenging in case of the HAD vehicles deployed in the open context. Moreover, the open context in general evolves over time and new phenomena emerges, even if the initial triggering conditions set was sufficient. For example, a decade ago e-scooters were not part of the road actors.

To the best knowledge of the author, the applicability of safety analysis techniques referenced by SOTIF have not been scrutinized for their ability to model performance limitations and triggering conditions. Considering the challenges discussed previously, novel solutions need to be introduced in case the traditional safety analyses technique cannot fulfil their modelling assumptions. Moreover, to perform SOTIF analysis for HAD vehicles operating in the open context, safety models that can incorporate randomness and lack of knowledge into the analysis needs to be introduced. Besides, to include the evolving nature of the open context, iterative augmentation of the safety analyses model may also be required based on the gathered knowledge about the context.

1.2 Thesis Goals

The goal of this thesis is to assess the existing safety analysis approaches advocated for the implementation of HAD vehicles' SOTIF analysis. Moreover, the thesis also intends to provide novel safety analyses techniques based on the modelling limitations present in some existing techniques. To address these issues, the following research questions have been formulated.

Research Question 1: Can existing safety analysis techniques such as FTA/ FMEA model SOTIF/ safety of the automated driving? If not, what aspects can they not model?

The aim of this research question is to critically analyse the applicability of the existing safety analyses techniques in the automotive domain to the HAD functions to assure SOTIF. The complexity and open context may result in uncertain behaviour. The safety analyses techniques include FTA, FMEA and STPA. This work analyses the limitations in fulfilling the assumptions when the existing safety analyses techniques are applied to the HAD functions to assure safety.

Research Question 2: Which safety analysis models are suitable to represent different types/facets of uncertainties encountered in complex systems and open context?

As discussed in the previous section, HAD vehicles are complex systems operating in the open context. The aim of this research question is to propose safety analyses models that can represent multi-faceted uncertainties encountered in a complex system and the open context.

Research Question 3: How can safety analysis models be applied to support an iterative augmentation of the safety analysis and enable discovery of new knowledge encountered in complex systems and open context?

Traditionally, safety analysis techniques are expert oriented and result in static models. SOTIF requires modelling of all the relevant triggering conditions to assess performance limitation. Identification of novel triggering conditions becomes specifically important and challenging in the case of the HAD vehicles deployed in the open context. The research question aims to explore hybrid (expert and data input) iterative methodologies in which novel triggering conditions can be discovered and identified systematically.

1.3 Scientific Contributions

In assessing the research questions defined in the previous section (Sec. 1.2), this dissertation makes the following scientific contributions.

Scientific Contribution 1: A thorough analysis of the safety models in the automotive domain concerning the safety of HAD vehicles to assure SOTIF.

This dissertation provides a detailed argumentation on the application of safety analyses techniques including FTA, FMEA and STPA on the SOTIF and safety of automated driving in general. First, the challenges in the safety assessment considering the nature of HAD vehicles as complex systems and open context nature of their deployed environment are discussed. Propositions about the important aspect of modelling are derived. The propositions are then assessed against the assumptions of the safety analyses techniques mentioned above.

Scientific Contribution 2: A categorization of uncertainties encountered in complex systems and open context to achieve SOTIF.

In order to address SOTIF, a novel categorization of uncertainties encountered in complex systems and open context is proposed. The categorization is done into aleatory, epistemic and ontological uncertainty. Part of this work has also been published [41].

Scientific Contribution 3: Provision of safety analysis model to represent uncertainties encountered in complex systems and open context to achieve SOTIF.

To build upon the uncertainty's categorizations, safety analysis models are introduced with the capability to represent uncertainties. First Bayesian Network (BN) is introduced to represent aleatory uncertainty. BNs are then extended by combining Dempster and Shafer Theory [104] concepts resulting in Evidential Network (EN) [109] and Extended Evidential Network (EEN) [3]. Epistemic uncertainty in

case of EN as well as both epistemic and ontological uncertainty in case of EEN can be represented in addition to aleatory uncertainty. An example constituting the assessment for SOTIF is also provided. Part of this work has also been published [3].

Scientific Contribution 4: Framework to provide an estimation and plausibilization of triggering conditions and the systematic discovery of new knowledge by identifying novel triggering conditions.

A framework is proposed in this dissertation that addresses two major aspects.

1. Estimation and Plausibilization: This aspect identifies relative frequency of performance limitations as well as the triggering conditions that affect the performance limitation by quantifying the underlying causal relations. Identification of triggering conditions assists in the identification of manageable set of triggering conditions, derivation of scenarios for V&V and open context model from the SOTIF standpoint. The identification also provides guidelines for SOTIF improvement measures.

2. Discovery of New Knowledge: This aspect of the framework provides a systematic iterative method to introduce new knowledge by incorporating human expert and data inputs in safety models.

The framework provides the two aspects while also keeping intact the uncertainties categorization and representation of the safety models. In order to support the framework implementation, an algorithm for each aspect of the framework is also proposed. Part of this work has already been published [1, 2].

1.4 Validation of the Approach

The contributions developed in this dissertation have been validated through a HAD vehicle demonstration case study.

In Ch. 4, a hypothetical example concerning the LIDAR based detection of surrounding vehicles is provided that validates the modelling abilities of BN, EN and EEN to model aleatory, epistemic and ontological uncertainty.

Moreover, a real-world example in which LIDAR based perception system to detect cars on the highway is used to support the implementation of the framework (Ch. 7). The implementation results in estimation and plausibilization of triggering conditions by quantifying their causal effect on performance limitation. SOTIF improvement measures are introduced to address the effect of relevant triggering conditions. Localized refinements are suggested where no certain decision can be made. This is followed by an iteration of the discovery of new knowledge through datasets. The identified novel triggering conditions are passed through a second estimation and plausibilization process.

1.5 Thesis Structure

After the introductory chapter, the fundamentals of safety engineering are explained in more detail in chapter 2. Moreover, a comprehensive insight into the related work is given. A special focus is on the safety standard ISO 21448 [54], as it is the most relevant standardization to this dissertation.

Chapter 3 provides an overview of challenges in the safety assessment and assurance of HAD vehicles. A critical argumentation of the safety analyses techniques

referenced by SOTIF is presented. The challenges in the safety assessment and assurance of HAD vehicles are summarized into distinct propositions.

Chapter 4 introduces the uncertainty categorization for SOTIF. Furthermore, safety analysis models to represent those uncertainties are provided. An example of SOTIF analysis is also provided in this chapter.

Chapter 5 proposes the novel framework to model, identify and discover perception performance limiting triggering conditions in automated driving. The framework is supported by two separate algorithms for estimation and plausibilization of triggering conditions and discovery of novel triggering conditions, respectively.

Chapter 6 and chapter 7 focus on the validation of the framework. Chapter 6 explains the real-world case study involved in validation, while chapter 7 presents and discusses the results of the implementation of the framework algorithms to the case study.

Chapter 8 summarizes the limitations and threats to the validity to the approaches introduced in this dissertation. The limitations are provided for the theoretical, methodological and implementable level aspects of the approaches.

Finally, chapter 9 provides conclusion and highlights the contributions to the stated research questions. Further research directions are also discussed in this chapter.

2

Safety Engineering Fundamental and Related Work

"That is part of the beauty of all literature. You discover that your longings are universal longings, that you're not lonely and isolated from anyone. You belong."
– F. Scott Fitzgerald, *American Novelist*

This chapter represents the safety engineering fundamentals and related works considered for this thesis. In Sec. 2.1, the fundamentals of safety engineering are presented. This includes the fault, error and failure model by Laprie et al. (Sec. 2.1.1). In Sec. 2.2, traditional safety analyses techniques are discussed in detail including FTA, FMEA and STPA. In Sec. 2.3, an overview of the relevant standardization activities is given. In Sec. 2.4, the search strings used to find the relevant publications are discussed and a detailed discussion on the salient approaches from the literature is presented in Sec. 2.5. Lastly, in Sec. 2.6 the chapter summary is provided along with a table summarizing the literature most relevant to this dissertation (Tab. 2.6).

2.1 Safety Engineering

Safety engineering is a discipline that involves assessment of hazardous situations and setting the tolerable frequency of those situations so that the system should be considered as sufficiently safe [114]. Safety engineering is applied through the product development lifecycle, to assess and reduce the risk that may lead to harm. Multiple safety engineering fields can be categorized. For example, functional safety focuses on the identification, analysis and prevention of failures of the E/E components [55] while SOTIF focuses on the intended functionality safety [54].

The safety engineering life cycle is aligned with the product development lifecycle. Product development life cycle (and consequently safety engineering lifecycle of a product) is governed by a breakdown of project activities. ISO 26262 "Road vehicles – Functional safety" addresses the breakdown of project activities through a V-model [55].

ISO 26262 covers the failure of E/E components and provides work product activities around this premise. With the emergence of automation functions, a novel safety critical phenomenon has emerged. This phenomenon relates to the absence of unreasonable risk due to hazards caused by functional insufficiencies and is termed SOTIF [54].

This dissertation focuses particularly on SOTIF and the safety of automated driving in general. These topics are not covered by ISO 26262, however, owing to completeness of the fundamentals of safety in the automotive industry, it is imperative to provide an overview. Faults become the basis for failures. ISO 26262 covers the

malfunctions and failures of E/E components of the automotive vehicle. To this end, the fault, error and failure description are introduced.

2.1.1 Fault, Error and Failure

Laprie et al. published a taxonomy for dependable and secure computing [9]. The work also provides a detailed taxonomy of fault, error and failures as well as the propagation connection between them. Here, the original definitions are provided followed by a detailed discussion on the taxonomy of each.

Definition 4 *Fault: The adjudged or hypothesized cause of an error is called a fault [9].*

Faults are hypothesized causes of errors and root causes of failures. Faults can be seen as an internal or external state of a system. Moreover, a prior presence of vulnerability in the system is necessary to enable external faults to harm the system.

Definition 5 *Error: The part of the total state of the system that may lead to its subsequent service failure is defined as error [9].*

System behaviour is defined as the sequence of states. In its essence, error then can be also defined as a cause of deviation from defined system behaviour.

Definition 6 *Failure: A service failure, often abbreviated to failure, is an event that occurs when the delivered service deviates from correct service [9].*

Laprie et al. argue that failures are caused by wrong functional specification or incorrect emanation of behaviour by a system not complying with the functional specification. The former is a product of development discrepancies, while the latter is either manifestation of development discrepancies or emergent attributes observed during operation. Transition from correct service to an incorrect service is considered as a failure. Moreover, it is only observed as it reaches the boundary of a system [9].

2.1.1.1 The Pathology of Failure: Relationship between Faults, Errors, and Failures

The generation and propagation of faults, error and failures is shown in Fig. 2.1 and the salient aspects of their propagation are summarized as follows.

1. Errors are produced only when a fault is active. At all other times, when errors are not being produced, faults are considered dormant.

2. A fault is activated by a specific activation pattern applied to a component.

3. Internal propagation of errors occurs within a component. External error propagation occurs when component B receives services from component A and through the internal error propagation, an error of component A reaches the service interface of component A.

4. Deviation from the correctly delivered service is caused by the propagated error that reaches the service interface and in return causes a service failure.

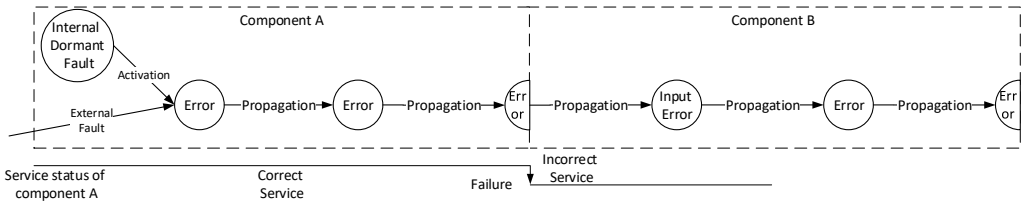


FIGURE 2.1: Error propagation as well as fault, error and failure propagation relationship. Computation processes cause internal propagation of errors. A failure is caused if the delivered service deviates from the correct service and reaches at the service interface [9].

Safety engineering necessitates the evaluation of faults and their causal factors that can lead to failures and consequently to hazards in safety critical systems. This evaluation is performed using safety analysis techniques.

2.2 Safety Analysis

Several safety analyses approaches are available across different domains in the literature [48]. Safety analyses techniques are used to identify the potential causes of hazards. These methods can be categorized as inductive (cause to effect) or deductive (effect to cause) approaches. They can also be categorized as qualitative and quantitative approaches.

In the automotive industry, FTA and FMEA are the most established safety analysis methods. FMEA as an inductive while FTA is categorized as deductive approach. In order to analyse a complex system and its environment, new safety analysis methodologies have emerged. These methods include STPA [70].

In the subsequent section, a summary of these methods is given, underpinning the basic definitions and descriptions.

2.2.1 Fault Tree Analysis

FTA is a safety analysis technique, in which an undesired state of the system is specified as a top event. Here, the system is analysed in the context of its operating environment to address all possible ways in which the undesired event can occur [129]. FTA uses a graphical notation similar to the Boolean logic using AND and OR logic gates to model causal events. An FTA is composed of several different symbols and events [129].

2.2.1.1 Primary Event

A primary event is an event in FTA that is not developed further. Probabilities for these events should be provided in case of quantitative FTA. There are four types of primary events, as follows.

Basic Event

A basic event is an event that requires no further development Fig. 2.2(a).

Undeveloped Event

An undeveloped event is an event that is not further developed for reasons that may include insufficient consequences or unavailable information.

Conditioning Event

A conditioning event is an event that provides restrictions and conditions that can be applied to any logic gate.

Intermediate Event

An intermediate event is an event that occurs through antecedent causes passing through gates.

2.2.1.2 Gates

There are two basic gates for FTA: AND and OR gate. All other gates can be summarized as particular cases of these gates.

OR Gate

The OR gate represents that output occurs when only one or more than one inputs occur. It is worth mentioning that *causality cannot pass through OR gate*. The OR gate can be quantified using the following equation.

$$P(X\ or\ Y) = P(X) + P(Y) - P(X \cap Y) \tag{2.1}$$

Two events can be mutually exclusive, independent or completely dependent. These properties for events modify Eq. 2.1 as follows.

$$\begin{aligned} &\textit{Mutually Exclusive} := P(X) + P(Y) \\ &\textit{Independent} := P(X) + P(Y) - P(X).P(Y) \\ &\textit{Completely Dependent (Y on X)} := P(Y) \end{aligned} \tag{2.2}$$

Since $P(X) + P(Y)$ always results in upper bounds, it is generally taken as an approximation. It is known as "rare event approximation" [129].

AND Gate

The AND gate represents that output occurs when all the inputs occur. Unlike OR gate, AND gate specify a causal relation between input and output. The AND gate can be quantified using the following equation.

$$P(X\ and\ Y) = P(X \cap Y) \tag{2.3}$$

Similar to OR gate, three equations can be defined for AND gates based on the relationship of the two events.

$$\begin{aligned} &\textit{Mutually Exclusive} := 0 \\ &\textit{Independent} := P(X).P(Y) \\ &\textit{Completely Dependent (Y on X)} := P(X) \end{aligned} \tag{2.4}$$

Eq. 2.1 and 2.3 represent the simple mathematical relation that can be used to propagate and evaluate the failure rate of the undesired event.

FTA provides a simple and yet powerful modelling method. The simple structure, straightforward mathematical principles and clear graphical notations have enabled the widespread usage of FTA in the industry.

However, owing to the simple modelling technique, FTA lacks in modelling the complex causal relations and variability [7]. To address this limitation, variation in FTA models with newer gate types have been introduced [129].

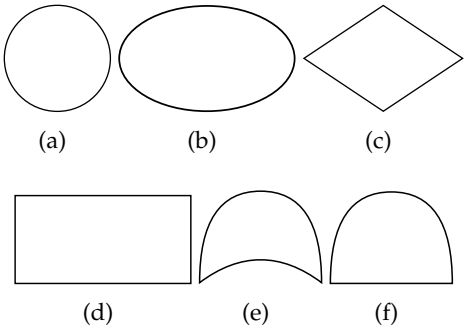


FIGURE 2.2: FTA most widely used symbols. **(a)** Basic Event **(b)** Conditioning Event **(c)** Undeveloped Event **(d)** Intermediate Event **(e)** OR Gate **(f)** AND Gate.

2.2.1.3 FTA Example

In order to further explain FTA, an example is provided in this section. For this example, the following simple system is considered.

"An emergency system comprises of two controllers: *A* and *B*. The controllers are powered by a single power supply. The controllers receive input from separate sensors *A* and *B*. At least one controller should provide an output to the emergency signal for the system to function properly. Random failures of the controllers are assumed to be negligible. Moreover, basic events are considered independent, unless stated otherwise. A schematic of this system is shown in Fig. 2.3."

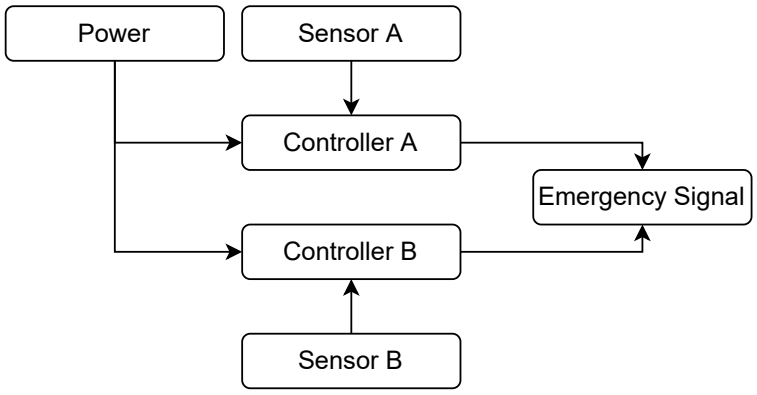


FIGURE 2.3: An emergency system used to provide modelling examples for FTA and FMEA techniques discussed in this chapter.

The FTA resulting for the emergency system described above (Fig. 2.3) is shown in Fig. 2.4. Evidently, the emergency system fails if both sensors (*A* and *B*) fail or the

power failure occurs. The probability of failure of the emergency signal is as follows.

$$P_{emergency\ signal} = P_{sensor\ A}.P_{sensor\ B} + P_{power} - P_{sensor\ A}.P_{sensor\ B}.P_{power} \tag{2.5}$$

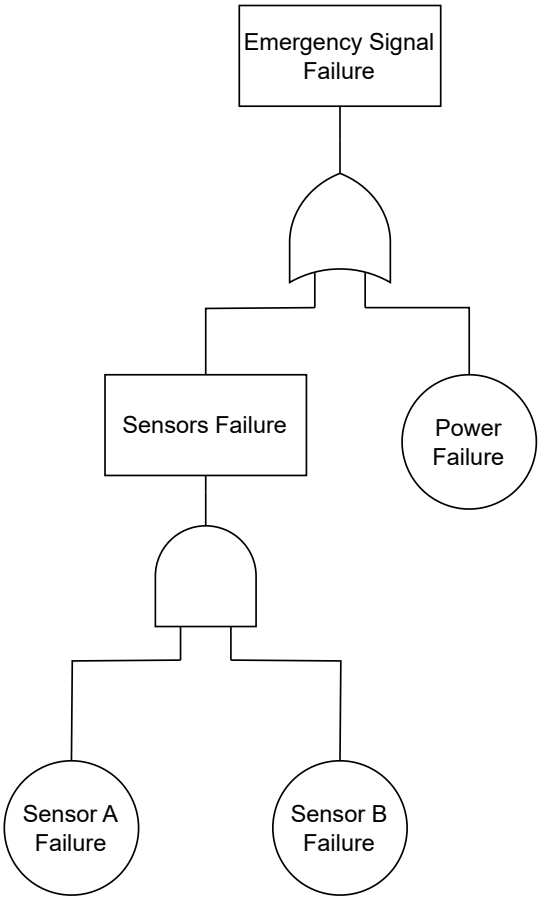


FIGURE 2.4: FTA resulting for the emergency system shown in Fig. 2.3. The system fails if either the power fails or both sensor A and sensor B fail simultaneously.

Eq. 2.5 summarizes the probability calculation for the emergency signal failure. It is important to note that independent events are considered in the example and the independent event relation is selected from Eqs. 2.2 and Eqs. 2.4.

2.2.2 Failure Mode and Effect Analysis

FMEA is a safety analysis technique used to define, identify and eliminate potential failures, problems or errors that may lead to a hazard. FMEA can be performed on the product designs, processes, systems and sub-systems.

FMEA identifies potential failure modes, their cause and effect as well as prioritize the failure modes through Risk Priority Number (RPN). FMEAs enhance safety and prevent defects. FMEAs are ideally conducted during product design or process development phases [77, 117].

FMEA starts with failure modes and their causal reasoning. RPNs are calculated using occurrences, severity and controllability of the failure modes. In this manner, every failure mode can be identified, ranked and if necessary, mitigated through mitigation measures. McDermott et al. [77] defines FMEA in ten steps (Tab. 2.1).

TABLE 2.1: FMEA steps as described by McDermott et al. [77].

Step 1	Review the process or product
Step 2	Brainstorm potential failure modes
Step 3	List potential effects of each failure mode
Step 4	Assign a severity ranking for each effect
Step 5	Assign an occurrence ranking for each failure mode
Step 6	Assign a detection ranking for each failure mode and/or effect
Step 7	Calculate the RPN for each effect
Step 8	Prioritize the failure modes for action
Step 9	Take action to eliminate or reduce the high-risk failure modes
Step 10	Calculate the resulting RPN as the failure modes are reduced or eliminated

2.2.2.1 Review the Process or Product

In this step, a blueprint of the product, system or subsystem is reviewed. In case of process, the relevant flowchart can be reviewed. This information provides the relevant understanding of the product, process, system or subsystem.

2.2.2.2 Brainstorm Potential Failure Modes

In this step, the potential failure modes are identified. The identification is generally based on the expert knowledge. Moreover, single point failures are generally considered in this step.

2.2.2.3 List Potential Effects for Each Failure Mode

Once the failure modes are listed, each failure mode is reviewed and potential effects of each failure are identified. Failure modes may have single or multiple effects. This step is important as it serves as the input to risk ranking assignment to the failures. This step can be seen as if-then process i.e., "if the failure occurs, then what are the consequences?"

2.2.2.4 Assigning Severity, Occurrence, and Detection Rankings

Each failure mode is assigned a severity, occurrence and detection ranking. The ranking varies at a 10-point scale, with 1 being the lowest and 10 being the highest rank. In this step a clear description of the points is required so that the ranking is performed consistently. Generic scale descriptions are present in the literature to support this process [77].

2.2.2.5 Assign a Severity Ranking for Each Effect

The severity ranking assesses how severe the effect will be, given the failure occurs. Rankings can be based on data analytics, experiences or expert judgements. Ranking is provided for each effect. Therefore, an effect has its own severity ranking given the failure mode occurs.

2.2.2.6 Assign an Occurrence Ranking for Each Failure Mode

The occurrence measures how often or probable the failure mode is? One of the best methods to assess the occurrence is arguably through actual data e.g., through failure logs or capability data [77]. If data is not available, the expert assesses the occurrence of the failure mode. Understanding the causes of failure may assist in this process. Occurrence can be measured based on the frequency or duration of the failure mode [55].

2.2.2.7 Assign a Detection Ranking for Each Failure Mode and/or Effect

The detection assesses the likelihood of detecting a failure mode. If the likelihood is low, the detection chance will be low as well. In some variants of the FMEA, a controllability factor is assessed instead of detection [55].

Controllability assesses the likelihood of containing the failure e.g., in case of a failure leading to a vehicle level hazard, how likely is it that the driver will control the vehicle?

2.2.2.8 Calculate the Risk Priority Number for Each Effect

The RPN is simply a product of severity, occurrence and detection/ controllability factors.

$$RPN = Severity * Occurrence * Detection(Controllability) \tag{2.6}$$

2.2.2.9 Prioritize the Failure Modes for Action

The failure modes can now be ranked based on the RPN, from highest to lowest.

2.2.2.10 Take Action to Eliminate or Reduce the High-risk Failure Modes

In this step, actions are devised to reduce the risk. Ideally, failure modes should be fully eliminated. If a failure mode is completely eliminated, the occurrence scale reaches zero, thus making RPN value zero, consequently. However, in some cases elimination of failure mode may not be fully achievable. In such cases, the severity scale reduces while detectability/ controllability scale is increased.

2.2.2.11 Calculate the Resulting RPN as the Failure Modes are Reduced or Eliminated

Once the action has been taken as described in the previous step the RPN is calculated again. A significant reduction in the RPN is expected if the action taken reduces severity, occurrence, detectability or their combination.

2.2.2.12 FMEA Example

Tab. 2.2 provides an implementation example of FMEA for power supply as a component of emergency signalling system as described in Fig. 2.3. The example considers transistor failure that can result in power failure. The severity, occurrence and detectability is calculated to determine the RPN.

TABLE 2.2: FMEA example for the system described in Fig. 2.3.

Component	Potential Failure Mode	Potential Effect of Failure	Severity	Potential Cause of Failure	Occurrence	Current Controls & Detection	Detection	RPN	Recommended Action	Target Completion	Action Taken	New RPN
Power Supply	Mosfet Failure	Emergency Signal Not Provided	8	...	6	None	10	480	Parallel Power	...	...	...

2.2.3 System Theoretic Process Analysis

System-Theoretic Process Analysis (STPA) is a relatively new hazard analysis technique based on an extended model of accident causation; System-Theoretic Accident Model and Processes (STAMP) [86]. In addition to component failures, STPA assumes that accidents can also be caused by unsafe interactions of system components, none of which may have failed.

There are four basic steps involved in STPA. In the following, each step is briefly described.

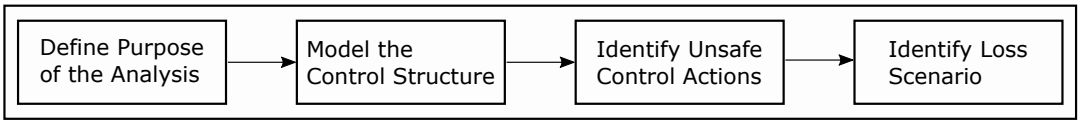


FIGURE 2.5: System Theoretic Process Analysis (STPA) [70] overview. STPA is a hazard analysis technique based on system engineering principles. It is based on the control loss accident causation assumption [71].

2.2.3.1 Defining Purpose of the Analysis

As shown in Fig. 2.5, the first step in applying STPA is defining the purpose of the analysis. Defining the purpose of the analysis has four parts.

- Identify losses
- Identify system-level hazards
- Identify system-level constraints
- Refine hazards (optional)

2.2.3.2 Model the Control Structure

In this step a hierarchical control structure is modelled. A control structure is a system model which comprises of control loops and feedbacks. The structure is hierarchical by nature i.e., order of control flows from top to bottom. Problems can occur at any point in the structure that can lead to Unsafe Control Actions (UCAs). Using the modelled control structures, UCAs leading to hazards are identified.

2.2.3.3 Identify Unsafe Control Actions

The third step is to analyse control actions in the control structure to examine how they could lead to the losses defined in the first step. The UCAs are used to create functional requirements and constraints for the system.

2.2.3.4 Identify Loss Scenario

The fourth step identifies the reasons why unsafe control might occur in the system. Scenarios are identified to explain the following:

1. How does the incorrect feedback, inadequate requirements, design errors, component failures, and other factors could cause UCAs and ultimately lead to losses?

2. How does safe control action, which might be provided but not followed or executed properly, lead to a loss?

A loss scenario describes the causal factors that can lead to the UCAs and to hazards. Though STPA provides a possible reasoning on the causal factors, it does not include a model to represent complex causal relations.

2.3 Standardization Activities

In safety engineering, standardization activities are performed to provide guidelines for safety assurance of products, activities and processes. There are many safety standards relevant to the automotive industry [123, 17, 56]. Two standards are discussed here based on their foundational guidelines for automotive industry and their relevancy to this dissertation.

2.3.1 ISO 26262

ISO 26262, "Road vehicles - Functional safety" is an ISO standard for E/E systems' *functional safety* (Def. 1). The E/E systems are installed on production road vehicles.

The standard is applicable to automotive development phase, ranging from the specification, conceptual design, integration, implementation, verification, validation and product release. The standard provides safety relevant product development guidelines at system, hardware and software level.

The goals of ISO 26262 are as follows.

1. Provision of a safety lifecycle for automotive products. This includes management, development, production, operation and decommissioning
2. Coverage of functional safety aspects of development process
3. Provision of risk classes for automotive products (Automotive Safety Integrity Levels, ASILs)
4. Derivation of safety requirements based on the ASILs
5. Derivation of validation requirements for safety assurance

Part 9 of ISO 26262 defines scope of safety analyses as V&V of safety concepts, identification of conditions and causes that can lead to a hazard as well as identification of safety requirements. The standard suggests usage of both qualitative and quantitative safety analyses for functional safety. In this regard, FMEA, FTA, ETA and HAZOP are considered as modelling techniques [55].

2.3.2 ISO 21448

ISO 21448, "Road vehicles - Safety of the intended functionality" is an ISO standard for E/E systems' *SOTIF* (Def. 2). SOTIF provides guidelines to reduce the unreasonable risk caused by:

- The insufficiencies of specifications emanating from the intended functionality.
- The insufficiencies of the specifications or performance limitations emanating from the implementation of E/E elements in the system.

Specific conditions can trigger the hazards that may lead to potentially hazardous behaviour [54]. The guidelines of SOTIF are applied to HAD functions where proper situational awareness is vital to safety. This means that SOTIF is intended to address the requirements for the HAD vehicle perception system used to perceive and understand the surrounding environment.

In order to achieve a sufficient level of SOTIF, the standard describes guidelines and activities. The activities initiate with the definition, specification and design of the function. Then identification of the performance limitations as well as the risk is evaluated. Functional insufficiencies and triggering conditions (e.g., environmental conditions causing miss-detection of certain objects or driver misuse) are identified, if the risk evaluated is unacceptable. The first identification step only considers the hazardous behaviour, while the subsequent step discusses the causes of the hazardous behaviour through identification of functional insufficiencies and triggering conditions. The system design is improved as deemed necessary through functional modification to reduce SOTIF risk. V&V are provided to prove the appropriateness of the design against known and unknown hazardous scenarios. SOTIF defines these activities in the following clauses.

- Functionality specifications, system design and architecture (see Clause 5).
- Identification and evaluation of the SOTIF oriented hazardous behaviour (see Clause 6).
- Identification and evaluation of the causes of the hazardous behaviour (e.g., by Cause Tree Analysis (CTA)) and functional insufficiencies, triggering conditions and to include sensing and planning algorithms (e.g., by Inductive SOTIF analysis) (see Clause 7).
- Improvement in the system design through functional modifications to reduce SOTIF risk (see Clause 8).
- V&V of the design appropriations with respect to the SOTIF (see Clauses 9-11).

Clause 7 provides methods for identification and evaluation of potential functional insufficiencies and triggering conditions. Safety analyses techniques discussed for this identification and evaluation are CTA, SOTIF FMEA and STPA. These safety analyses techniques are adaptation of FTA, FMEA and STPA discussed in the previous sections.

ISO 21448 is deemed as the most relevant standardization to this dissertation. The standard along with its proposed guidelines and safety analysis techniques is revisited in the next chapter to provide a more in-depth and critical review on the topic. This discussion is out of the scope of this chapter.

2.4 Assessment of Related Work

The assessment of the state of the art for this dissertation is performed by keyword searches and manual snowballing. In general, publications related to SOTIF, causal factor and triggering conditions are selected. The selected publications are then passed through a snowballing search.

Tab. 2.3 shows the different strings used for keyword search and resulting corresponding number of publications. In total, **658** publications resulted out of this step. Fig. 2.6 shows the flow of the publication selection process. The first filter applied is the duplication and language in which the publication is written, resulting

in **207** publications. Based on the title and abstract of the remaining publications, **69** publications are selected. For these publications, a further reduction is performed based on the content of the complete manuscripts which results in **35** publications. Finally, **21** publications are selected from this process. This selection is based on the relevancy of the publication to this thesis. Manual snowballing results in **6** publications which are also assessed in detail. The search is conducted on the IEEE, Google Scholar, Springer and Science Direct. Moreover, the search was performed on publications available before 1st January 2022.

TABLE 2.3: Search strings used to extract publications related to this dissertation. In total, the search string resulted in 658 publications. Duplication and language based refinement resulted in 207 publications.

Search String	Number of Publications
"SOTIF" AND "21448" AND "causal"	78
"SOTIF" AND "21448" AND "triggering"	123
"SOTIF" AND "automated" AND "causal"	94
"SOTIF" AND "automated" AND "triggering"	137
"SOTIF" AND "autonomous" AND "causal"	91
"SOTIF" AND "autonomous" AND "triggering"	135

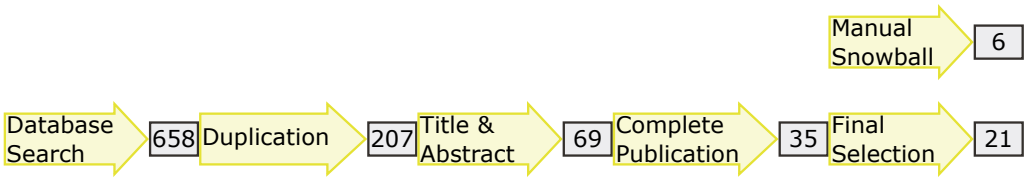


FIGURE 2.6: Publication selection process adopted in this dissertation. Each step of the process reduces the amount of the publication considered while increasing the specificity of the scope and relevancy of the publication to the thesis. In total, **27** publications are read and assessed in detail.

2.5 Salient State of the Art Approaches

In this section, some of the most important research related to this thesis is presented. Most of the work represented here comes from recent and manually selected publications emphasizing the novel nature of the problem addressed in this thesis.

2.5.1 Hazards Identification Originating from Variability

Ali et al. [7] analyse the hazards arising due to variability, complexities and uncertainties in collaborative Cyber Physical Systems (CPSs). Environmental, infrastructural, spatial and temporal variabilities are considered as factors causing uncertainties. They also develop a fault traceability graph to trace the faults considered by multiple hazard analyses in the collaborative CPSs with variability. The authors also extend the safety analysis techniques (FTA, FMEA and ETA) to explore hazards with variability. The motivation behind this research is to minimize known unsafe and unknown unsafe scenarios to achieve SOTIF by identifying associated risk with

variabilities in the environment, spatial and temporal estimations. FTA is extended by introducing variability in the OR and AND gate. Similarly, ETA and FMEA are extended with variable initiating event and variable point, respectively. In doing so, the extended models allow the multi-state variables inclusion, which increases the modelling capabilities of the studied techniques. However, this methodology does not provide comprehensive solutions on modelling the complex causal relations that HAD vehicle and the environment emanate.

2.5.2 Probabilistic Bounds on Hazards

Edward Schwalb [103] provides a probabilistic framework for incrementally bounding the residual risk associated with autonomous drivers and enabling the quantifying progress. The methodology provides probability calculation on the hazards that occur individually and in combinations.

The work introduces continuous monitoring by autonomous drivers for imminent hazards and selects actions that maximize the Time To Materialization (TTM) of these hazards. The approach also enables implementing the continuous expansion of SOTIF through measurement of improvements from regressions using posterior probabilities.

The study proposes to use FMEA and FTA to model limitations. The author also proposes to extend the implementation to all traffic participants. In order to calculate Time To Collision (TTC) metric, the author suggests using residual error inclusion through testing. However, based on different train and test data selection, this inclusion may result in hazardous scenario. In this way, this work partially addresses the causalities and their effects present in different datasets, even when residual risk is addressed through random errors.

2.5.3 Identification and Quantification of Hazardous Scenarios

Kramer et al. [67] provide integrated method for safety assessment of automated driving functions. This covers the aspects of functional safety and SOTIF, including identification and quantification of hazardous scenarios. They also provide a causal chain analysis technique to identify and model SOTIF related hazards.

The methodology starts with hazard identification and consequently identification of causal factors through causal chain analysis. The causal factors constitute the hazardous scenario, which can be assessed for risks. Based on the risks, the requirements can be defined. On the other hand, if the risk is considered tolerable, the V&V process can be initiated. Identification of functional insufficiencies, causal chain analysis and derivation of triggering hazardous scenario steps of the publication are relevant to this dissertation. Identification of functional insufficiencies is provided through FMEA oriented analysis while causal chain analysis is provided through an extended version of FTA (using inhibit gate). Hazard triggering scenarios are defined using traffic sequence chart. These scenarios are quantified using probability of occurrence, minimal cut-sets, error rate calculation and severity measurement. This technique, however, does not provide argumentation on the adequacy of the modelling techniques used.

2.5.4 Criticality Analysis for V&V

Neurohr et al. [80] propose a methodical criticality analysis that maps an infinite-dimensional domain onto a finite and manageable set of artefacts. These artefacts capture and explain the emergence of critical situations for automated vehicles. The

study proposes a combined approach of expert-based and data-driven methods i.e., a hybrid approach to identify relevant phenomena and explain the underlying causalities.

The methodology initiates with the identification of a critical phenomenon, followed by hypotheses over the causal explanation of the emergence of the phenomenon. These hypotheses are plausibilized. If the plausibilization is not possible either because it is not a valid hypothesis or because refinement is required, feedback towards initial steps is requested. If the plausibilization is possible, the phenomena are catalogued. The resulting phenomena can be used as a criticality measure for scenarios. However, this work provides a more theoretical view of the problem.

2.5.5 Uncertainty Treatment in Safety Analysis

Gansch [40] provides a system theoretic approach to incorporate diverse types of uncertainties (epistemic, aleatoric and ontological) into the safety analysis performed for automated driving functions. The work argues the existence of high degree of uncertainties in the performance of technologies involved in automated driving functions. It also claims that SOTIF intends to address and reduce the present uncertainties to residual risk.

The work provides a solution with the implementation using BN [88, 66], while also extending it with evidence theory [28, 104]. The provided solution does not consider inclusion of data in the construction of the BN. It does not address the modular extension of the BN when novel triggering conditions are identified in the identification process of unknown unsafe scenarios.

2.5.6 Safety Assessment of Environment Perception

Berk [14] provides safety assessment methods for the perception sensors to assure automated driving safety. The method develops reliability requirements for individual sensors. This is based on the stochastic description of reliability, conceptualization of sensor data fusion and statistical dependence models for sensor errors. This thesis also proposes an approach in which sensor perception reliability is learnt without a reference truth by exploiting sensor redundancy. The work is based on the reliability-based approach and does not address the problem from a SOTIF viewpoint.

2.5.7 Quantitative SOTIF Analysis

Wendorff [133] provides a methodology by quantifying safety performance of an automated driving system. The methodology is based on environmental, obstacle and vehicle model. The work utilizes the extension of FMEA approach quantifying the probability of a lack of safety performance and identifying corner cases for validation tests.

2.5.8 Scenario Oriented Safety Analysis

Another important approach that has gained importance in the safety analysis methods is based on scenario analysis [94]. Though, it can be argued that all previously mentioned approaches model scenarios in some specified settings as well, scenario-based safety analysis covers approaches with loosely connected ideas of

failure mode, faults and errors. This type of analysis is deemed closer to V&V methods [54, 64, 94]. For example, failure modes, their causes and undesired events together constitute unsafe scenarios partially. These approaches attempt to structure the deployed environment, identify edge cases while condensing the amount of scenario driven and addressing the safety assessment problem.

In our understanding, scenario-oriented safety analysis shrinks the gap among design-oriented safety analysis and V&V. The need for scenario-based approach stems from SOTIF [54] which attempts to address the closure of HAD vehicle safety from the known and unknown unsafe scenario identification and mitigation.

2.6 Summary

Tab. 2.6 summarizes the related work in the field of safety analysis implementation assessed and relevant to this dissertation. The table is first sorted by the "Year" of publication and then alphabetically on the "Author & Publication" column. The empty cells indicate missing information about the column in the publication. All other information provided is collected under the light of this thesis. It should not be misinterpreted as the only relevant information for a given column.

The column "Scope" describes the solution provided by the publication that is also relevant to this dissertation. The "Analysis Tool & Approach" shows the safety analysis implementation tools and analyses methodologies used in the publication. The "SOTIF Contribution" summarizes the clause of the standard [54] towards which the publication contributes. The last row represents the implementation of this dissertation.

Most of the publications involved stem from the automotive and automated driving industry. This can be related to the fact that the SOTIF standard to which this dissertation provides solution also stems from the same industry.

TABLE 2.4: Safety analysis approaches reviewed and assessed for this dissertation.

Author & Publication	Scope	Analysis Tool & Approach	SOTIF contribution	Year
Bai et al. [10]	Identification of external influencing factor	Analytical methods	Validation	2019
Berk [14]	Safety assessment methods for perception sensors	stochastic description of reliabilities and statistical dependencies models	Validation	2019
Poddey et al. [91]	Limitation & possible solution of validation approaches		Validation	2019
De Gelder et al. [42]	Assessment & quantification of risk for automated driving scenario	Expert causal model of scenario, monte carlo simulation for severity	V&V of scenarios	2019
Gansch [40]	SOTIF analysis	Probabilistic, Bayesian Network (BN)	analysis of triggering condition and performance limitations	2019
Kirovskii et al. [65]	Integration of functional safety [55] and SOTIF [54]	Triggering event analysis, acceptability of triggering conditions		2019
Vander et al. [127]	Hazard and triggering environmental conditions identification	Extended HAZOP and FTA	Hazard and triggering conditions analysis, triggering scenario identification	2019
Martin et al. [74]	Identification of triggering conditions	HAZOP, Expert causal analysis	Triggering conditions identification & analysis	2019
Schnellbach et al. [99]	Summarizes the ISO 21448 [54]			2019
Schwalb [103]	Bounds on hazards' risks	Probabilistic, extended FMEA & FTA	Verification of scenarios	2019
Wendhorff [133]	Quantification of the safety performance of HAD vehicles	Extend FMEA	Analysis of triggering conditions	2019

(To be continued)

Author & Publication	Scope	Analysis Tool & Approach	SOTIF contribution	Year
Becker et al. [13]	SOTIF implementation on TJP	STPA		2020
Khatun et al. [64]	Scenario based HARA	HAZOP	HARA for SOTIF	2020
Ali et al. [7]	Safety analysis extended to uncertainty & variability	Extended FTA, FMEA & ETA	Analysis of triggering conditions	2020
Kramer et al. [67]	Identification and quantification of hazardous scenarios	HAZOP & extended FTA	Identification of unsafe scenarios and triggering conditions	2020
Bannour et al. [11]	Designing & generation of logical scenario	Formal methods	Scenario generation for SOTIF V&V	2021
De Gelder et al. [27]	Data driven risk quantification	Probabilistic, Scenario based	Validation of scenarios	2021
Neurohr et al. [80]	Identification of critical phenomena & scenarios	Criticality	V&V	2021
Hussain et al. [53]	Enhanced Fault Traceability & Propagation Graph based safety analysis	Extend FTA, FMEA & ETA		2021
Scholtes et al. [100]	Identification of influencing factors for RADARs	expert knowledge, data driven, market based and 6-layer structure [101]	Identification of influencing factors imposing performance limitations	2021
Zhang et al. [135]	Hazard analysis for SOTIF,	STPA, Causation	Hazard analysis, identification of triggering conditions and performance limitations	2021
Khatun et al. [63]	Scenario based safety analysis		Identified scenarios & triggering conditions	2021
Zhang et el. [136]	Provision of literature on critical scenario identification			2021
Scholtes et al. [101]	Structuring the urban environment		Formalization of scenarios for V&V	2021

(To be continued)

Author & Publication	Scope	Analysis Tool & Approach	SOTIF contribution	Year
Thomas et al. [122]	Hybrid Causal Logic Framework	ESD, FTA, BN	Analysis of triggering conditions & performance limitation	2021
Wu et al. [134]	Modelling uncertainties of interaction for SOTIF	Complex network theory		
Burton [18]	Causal models for ML	Theoretical causal models, FMEA and FTA		2022
This thesis	**Identification & discovery of triggering conditions**	**Probabilistic models, BN**	**Identification of triggering condition and analysis, SOTIF improvement measures, V&V of scenarios**	**2023**

Conclusion that can be drawn from the salient state-of-the-art approaches (Sec. 2.5) and Tab. 2.6 are associated with the utilization of the legacy analysis methods (Sec. 2.2), rigorous involvement of probability theory as well as related modelling techniques, identification of triggering conditions, causal factors and critical scenarios. Moreover, some techniques also indicate the usage of hybrid approaches for analysis.

Some publications assessed, implement the HAZOP, FTA, FMEA and ETA with some tailored extensions. Such extensions have resulted in variability inclusion in the FTA, FMEA and ETA [7, 53], extended keywords list in HAZOP [127] and use of inhibit gates in the FTA [67]. Inhibit gate in FTA were introduced in the FTA handbook [129] and will be addressed in the subsequent chapters. Thomas et al. [122] also utilizes the three analyses (ESD[1], FTA, BN) to model various interactions. These techniques are expert driven; they solely rely on expert knowledge about the system and its environment related hazards and their causal factors. Identification of causal factors under this arrangement requires experience and this may become cumbersome especially in the case of automated driving, where resulting causal factors may be many and not apparent.

The second scope identified through the assessment of relevant approaches relates to identification of critical scenarios and probabilistic methods. Here, the two terms are intentionally discussed together to address identification of causal factors and performing causal analysis. Approaches related to scenarios such as risks assessment [27], criticality analysis for validation [80], scenario risk quantification [42] and scenario-based safety analysis [63] exercise a grey box implementation methodology. These techniques do not provide a detailed understanding of triggering conditions and improvements at design and development stage. Neurohr et al. [80], Gansch [40] and Schwalb [103] provide some insights on the causal factors' identifications. However, identification of novel triggering conditions still depends on the propositions provided experts, OEM's specifications and literature studies, to the best understanding of the author of this dissertation. These approaches are in general hybrid; they are both expert and data driven.

The last row of Tab. 2.6 represents the approach this dissertation studies. Presented approach is based on expert knowledge and data-driven engineering processes which uses the probability and causal theory as well as graphical models to represent the triggering conditions and performance limitations while also addressing the causal factor analysis. The approach models complex causal relations and can represent uncertainties. Systematic discovery of the novel triggering conditions is also provided. The resulting models can be used to identify and analyse causal factors for SOTIF improvement measures while it can be also used as input scenario catalogues for V&V.

[1] Event Sequence Diagram

3

Challenges in Assuring Safety for Automated Driving

"Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less."
– Marie Curie, *Polish/French Physicist and chemist*

In this chapter, a detailed description of the challenges of automated driving safety is provided. In Sec. 3.1, a HAD vehicle functional architecture based on *sense, plan* and *act* is provided. In Sec. 3.2 the description of HAD vehicle's automation levels is introduced as provided by Society of Automotive Engineers (SAE) in standard J3061 [97]. This is followed by a system theoretic discussion of the challenges to assure safety of the HAD vehicles operating in the open context (Sec. 3.3). In Sec. 3.4, a summary of ISO 21448 is provided. A synergy between the discussed challenges and analysis techniques is constructed through argumentation to address the limitations in the mentioned techniques in Sec. 3.5.

3.1 HAD Functional Architecture

The functional architecture of HAD vehicles is generally described by the sense, plan and act paradigm, the orthodox architecture used in robotics [31, 105, 19]. Fig. 3.1 represents such architecture for the HAD vehicles. Sensing block gathers information about the environment by using multiple and/or diverse perception sensors e.g., RADARs, LIDARs and cameras. This block may also include other information channels e.g., digital maps and V2V/V2X infrastructures for contextual information. The gathered information is fused and a vehicle environmental model is generated [31]. Based on this model, planning block interprets a driving situation

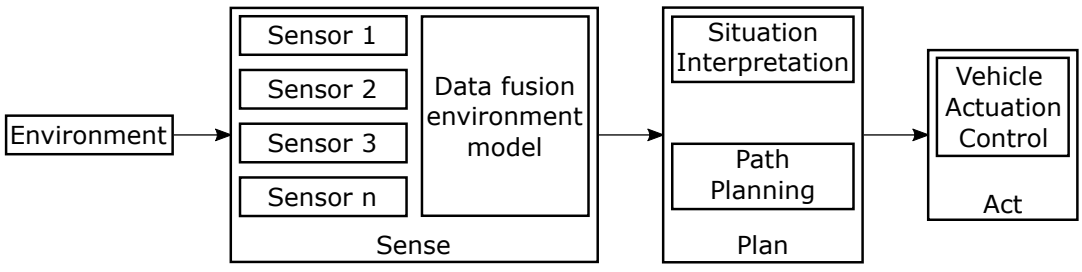


FIGURE 3.1: Sense, plan and act functional architecture of the HAD vehicle. Sensing block collects environmental information. The gathered information is fused and modelled. The model is used to assess driving situation and driving strategy is derived. Finally, act block implements the driving strategy. [19].

e.g., vehicles' as well as other traffic participants' position, velocity and trajectory. A trajectory along with driving parameters e.g., desired velocity and acceleration based on the initial transportation goal (drive A to B) and current situation is derived. Finally, the act block then executes the planned driving strategy using vehicle actuators (i.e., steer, engine and brakes). The functional architecture (Fig. 3.1) is a general description and may vary.

Sense, plan and act may fail or not provide the intended function. This may lead to a hazardous situation during operations. Moreover, HAD vehicles are intended to be deployed in the open context [19]. The open context may also emanate uncertain and unsafe behaviour due to the performance limitation and functional insufficiencies in sensing and interpreting the scenarios. This highlights the safety critical nature of sensing the open context, the planning of driving situation and the actuation.

3.2 SAE Level of Driving Automation

In order to fully grasp the challenges faced by HAD vehicles, categorization of HAD vehicles' automation capabilities are summarized. Categorization provided is based on the SAE taxonomy of driving automation [97]. Tab. 3.1 provides a summary of different levels of driving automation. For the exact definitions readers are referred to the SAE J3061 standard [97].

SAE Level	Description
L0	No Automation: No autonomy is present. Driver is responsible for all the dynamic driving tasks.
L1	Driver Assistance: Driver assistance features are present. The driver is responsible for all the dynamic driving tasks.
L2	Partial Automation: Combined driver assistance feature are present. The driver is responsible for all the dynamic driving tasks.
L3	Conditional Automation: Some automated driving modes are present. The driver monitors and takes control whenever required.
L4	High Automation: The vehicle performs the dynamic driving tasks under most conditions. The driver has the option to take control.
L5	Fully Autonomous: The vehicle performs all dynamic driving tasks under all conditions.

TABLE 3.1: SAE level of automation [97]: SAE International provides **6** level of driving automation, from no autonomy to full autonomous vehicle.

As shown in Tab. 3.1, there are two major shifts between the levels, i.e., from level 2 to level 3 and from level 3 to level 4. The shift between level 2 to level 3 lies in the fact that drivers are not responsible for driving from driving automation level 3 and above. However, in level 3, a driver might be requested to take control of the vehicle whenever requested by the automated driving features. In level 4 and level 5, the human driver is completely out of the driving task loop.

3.3 Challenges in Automated Driving Safety

The proliferation of HAD vehicles as the means of transport for masses is far from guaranteed. A HAD vehicle is a safety critical system which requires emanation of

reasonably safe behaviour during their operations. HAD vehicles bring new challenges when it comes to assure safety. In standard vehicles, a human driver is responsible for making the decision about driving manoeuvres based on the perceptual judgement of the environment and acting through the vehicle actuators. With HAD vehicles, the task of the human driver is taken over by the vehicle, partially or completely. While computer systems excel over humans at tasks like computation, they may find rather simpler intuitive tasks challenging such as recognizing novel scenarios. Moreover, complexity of these systems, infinite number of scenarios as well as causal factors and complex system interactions with its context brings in novel problems. A consensus on how to handle the safety for these technologies still lacks behind in literature. ISO 21448 [54] can be seen as one of the first attempts in this direction. In the subsequent sections, a system theoretic view of the challenges to assure safety of the automated driving is presented.

3.3.1 System Level Challenges

HAD vehicles are inherently complex systems. A complex system is composed of multiple components with the ability to interact with each other. Its behaviour is intrinsic in nature and difficult to model due to dependencies, relationships and interactions. Models that represent these systems while ignoring intricate system behaviour or characterizing it as a noise will be inaccurate [126]. Such systems manifest properties related to non-linearity, emergence, spontaneous order, adaptation and feedback loops [20, 34, 118, 120].

3.3.1.1 Non-linearity

HAD vehicles may behave non-linearly; they may respond to the inputs differently depending on their current state or context [20]. With techniques such as Deep Neural Networks (DNNs) used to perceive the environment [46], HAD vehicles may produce irregular outputs with small perturbations to the input e.g., adversarial examples [29].

3.3.1.2 Emergence

Another common feature related to complex systems and consequently HAD vehicles is the presence of emergent behaviour [70]. Emergent behaviour results when combination of parts result is unpredictable behaviour. These are the characteristics of the system that emerge from the dependencies, interactions or relationships between different components.

3.3.1.3 Semi-permeable Boundaries

The boundary between the system and its environment is described by the selection of the system of interest. Depending on the analysis objectives, this boundary may vary. For instance, if only HAD vehicles' functions are considered, the system consists of component for sensing the environment, planning the future states and providing actuation to achieve those states. However, if HAD vehicles are taken as mobility services, then other traffic participants, road infrastructures and environmental conditions should also be considered [85].

3.3.2 Open Context and its Challenges

The environment in which the system is deployed in is important for safety engineering as it regulates the function and performance of the system. For HAD vehicles, the deployed environment specification is of paramount importance e.g., this boundary definition is required for the Operational Design Domain (ODD) definition of SAE level of automation defined in Sec 3.2.

In the case of HAD vehicles, their deployed environment is essentially defined as open context [19, 91] which brings in its challenges from the safety standpoint. The challenges of such an environment are described through the following attributes.

- Unstructured: The environment in which HAD vehicles are deployed is unstructured. This indicates that there lies inherent randomness in how the participants of the environment behave. Even for SAE L1 or L2 vehicles (Tab. 3.1), where the system is simple and deterministic as well as the environment somewhat restricted, the ODD still cannot be defined deterministically. Provision of a structure to model the open context from the HAD safety aspects standpoint remains a challenge.

- Complex Interactions: Elements of the open context interact with each other and produce complex interactions. This results in causal factors present in the open context with complex causal relations. These factors can adversely affect the safety of HAD vehicles. For example, weather and road surface may interact to produce multitude of reflections from the road surface. They may also produce a *confounding phenomenon*, a spurious relation between the dependent and independent variable (Fig. 3.2 and Fig. 3.3). Elements of the open contexts also exhibit randomness in the causal relation, at least at the level of abstraction they are studied e.g., how different road surface conditions impact the reflections of road. Causal relation identification and explainability requires methods that can model this type of complexity.

- Evolving: Another important feature of the open context is its evolutionary nature. New traffic participants may need to be added in the relevant ODD. For example, delivery robots are slowly becoming part of the road infrastructure and may impact ODD in the future [12]. The evolving nature of the open context demands a framework that can incorporate new knowledge and facts (e.g., in terms of data, information) into the existing understanding of the open context. It can also provide indicators of the evolved nature of the open context, which can be further analysed.

- Unknown Elements: In the HAD vehicles' environment, instead of closed world assumption, where all the elements of the context are known, an open world assumption is taken, especially for SAE L4 and L5. This brings in the impact of unknown causal factors present in the environment but not considered when considering safety of the HAD vehicle. For example, for HAD vehicle's perception system operating in a specified environment, triggering conditions such as occlusion and truncation has been considered as part of analysis of SOTIF by calculating the effects on False Negative (FN) probability [1]. However, it is entirely possible that other elements of the context e.g., traffic density also affect the FN probability of the perception system [2].

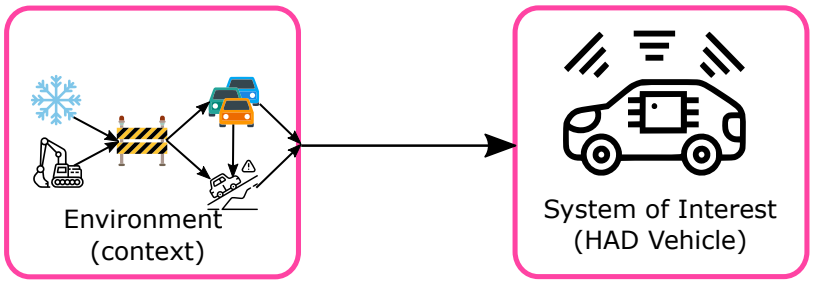


FIGURE 3.2: Complex causal relations present in the open context. Open context contains complex causal relations which cannot be modelled by traditional safety analysis techniques.

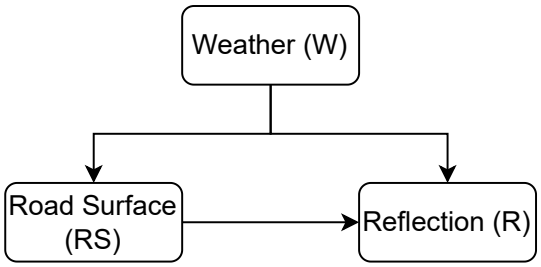


FIGURE 3.3: Example of a confounding phenomenon between environmental factors: Environmental causal factors may interact with the with one another and result in causal relation. Here, the *Weather (W)* nodes effects both *Road Surface (RS)* and *Reflection (R)*. Unobserved latent variables resulting from the *lack of knowledge* about the deployment environment may affect in the similar manner and affect the efficiency of safety analysis.

3.3.3 System and Context Interaction

HAD vehicles and how they interact with their environment play a critical role in the safety assessment. Especially, at the sense function block (Sec. 3.1), environmental conditions may impact the sensing performance (Fig. 3.4) e.g., RADAR sensor performance is susceptible to rainy weather [52]. A categorization on the properties of these interactions is discussed as follows.

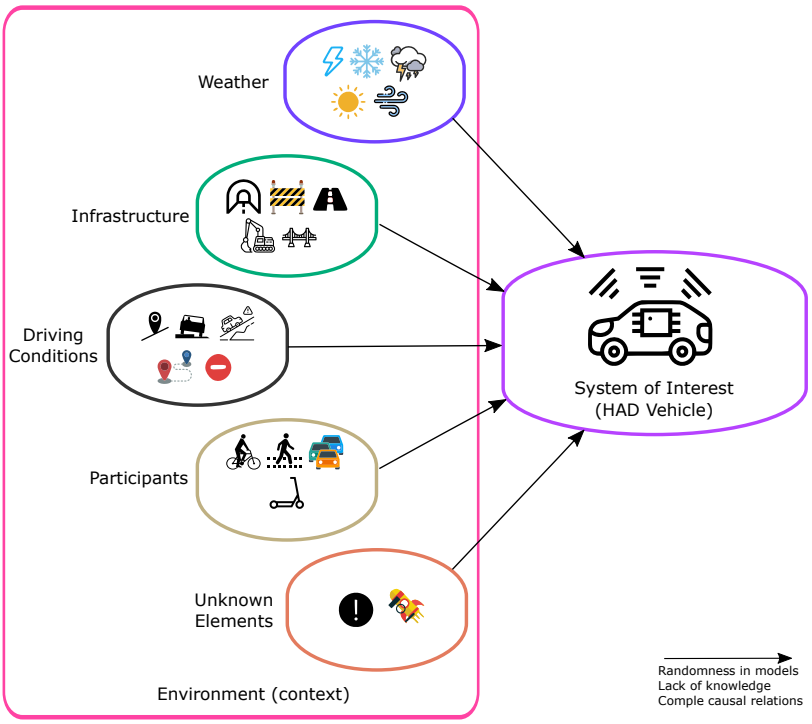


FIGURE 3.4: Causal effects of the environmental causal factors on the Highly Automated Driving (HAD) vehicle (system of interest). Due to the open context nature of the deployed environment, multiple causal factors affect the environmental perception sensors of the HAD vehicle. The causal effect of these factors is also random in nature at the abstraction level they are generally studied. Besides, the causal relations may also be complex in nature. Moreover, the lack of knowledge about existence and occurrence of the unknown causal factors is present for open context.

3.3.3.1 Causal Relations

Modelling causal interactions between system and context can be in-deterministic due to its complexity and non-linearity. For example, modelling the causal impact of rain on the camera sensor (sense functional block) deterministically is a challenging task. Rain characteristics e.g., droplet size, intensity of rain etc. will have varying impact on the camera sensor performance. Moreover, the possible confounding phenomenon (Fig. 3.5) demands a more detailed understanding of the causal relation argumentation.

Fig. 3.5 shows a hypothetical example of a *confounding* phenomena in which the FN probability of a camera sensor is evaluated to assess its performance. A camera-based perception system performance is known to be affected based on illumination [128, 79, 47, 44]. Moreover, both illumination and FN rate of perception systems are susceptible to weather conditions [79, 47, 44].

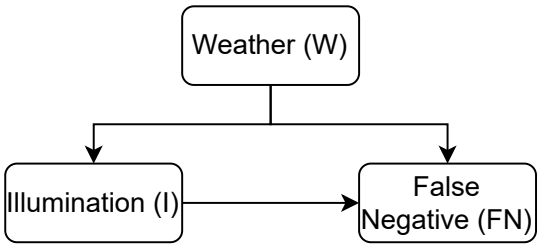


FIGURE 3.5: Example of a complex interaction between system and environment. Environmental causal factors may interact with the HAD vehicle situational awareness in a complex manner. Here, the *Weather (W)* nodes effects both *illumination (I)* and *False Negative (FN)* probability of the camera detection. Unobserved latent variables resulting from the *lack of knowledge* about the system and its deployment environment may affect in an analogous manner and effect the efficiency of safety analysis.

3.4 Safety of the Intended Functionality

3.4.1 Basic Architecture

SOTIF provides an analogous architecture to fault, error and failure described in Sec. 2.1.1.1. Triggering conditions and their combinations lead to functional insufficiencies which can result in hazardous behaviour. Functional insufficiencies may consist of insufficiencies of specifications or performance limitations (Fig. 3.6). For example, severe weather (triggering condition) can induce a miss-detection of road object (performance limitation) which then can lead to hazardous behaviour.

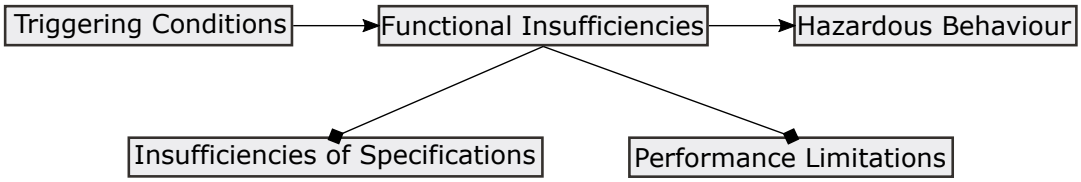


FIGURE 3.6: Basic Architecture of SOTIF. Triggering conditions and their combinations thereof, lead to functional insufficiencies which can result in hazardous behaviour. Functional insufficiencies may compose of insufficiencies of specifications or performance limitation.

3.4.2 SOTIF Activities

SOTIF provides a flowchart to summarise the activities (Fig. 3.7). The broad evaluation of the SOTIF is as follows.

- Evaluate by analysis
- Evaluate by V&V (evaluate known and unknown hazardous scenario)

Another important aspect of SOTIF is the *improvement measures* discussed in the clause 8 of the standard.

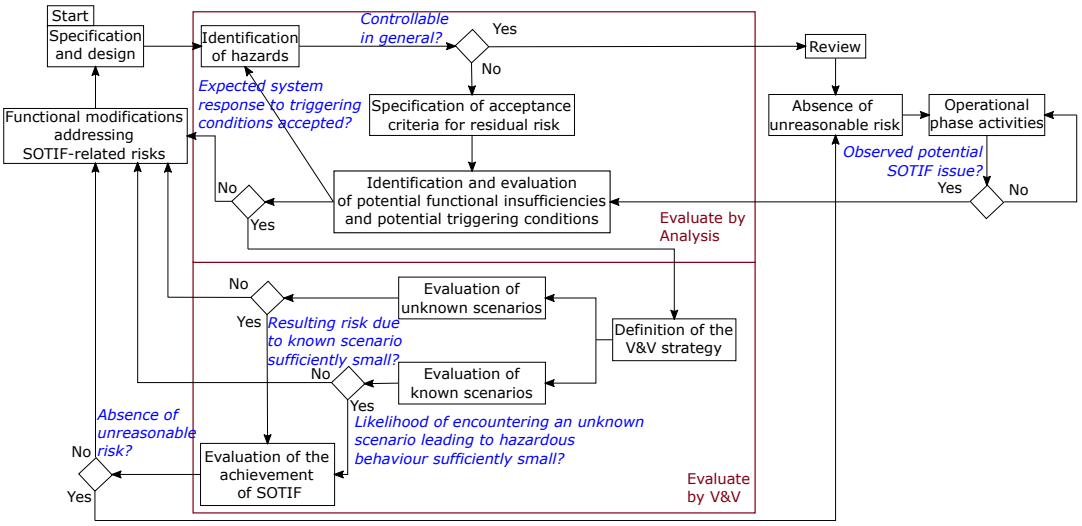


FIGURE 3.7: Flowchart of the ISO 21448 activities [54]. The flowchart summarizes the SOTIF activities into two main evaluations: (1) By analysis (2) By V&V. The evaluate by analysis provides input for V&V strategy.

3.4.2.1 Evaluate by Analysis

In the analysis block (Fig. 3.7), the scope of this thesis is limited to the identification of the causes of hazardous behaviour and functional insufficiencies pertinent to the perception and algorithm related to perception. The identification of performance limitations and triggering conditions can be performed quantitatively or qualitatively. In this regard, the standard refers to deductive and inductive methods including CTA, STPA and FMEA. The target of these analyses is to increase the understanding of the potential functional insufficiencies of the systems and support the identification of triggering conditions.

SOTIF argues that a systematic method can be established to perform the analysis of functional insufficiencies and triggering conditions. The provision of triggering conditions is governed by scenarios extracted from the ODD. In this sense, it can be inferred that the known/ unknown triggering conditions are part of the known and unknown hazardous scenarios.

3.4.2.2 Evaluation of Verification and Validation

Evaluation by V&V constitutes the evaluation of known and unknown unsafe scenarios (Fig. 3.7).

Evaluation of known scenarios constitute clause 10 of the standard [54]. This clause supports identification of potentially hazardous scenarios, evaluation and verification of the system's functionality in the known hazardous scenarios. The verification strategy is divided into sensing, planning, actuation and integrated system verification. Moreover, to provide residual risk acceptance, the risk of known hazardous scenarios should comply with the acceptance criteria and no known scenario should lead to unreasonable risk.

Evaluation of unknown scenarios constitutes clause 11 of the standard [54]. The clause demonstrates that the residual risk from the unknown hazardous scenarios meets the acceptance criteria. Unknown scenarios originate from reality. Methods to evaluate the residual risk include validation of robustness, randomized input tests, vehicle level testing of edge and corner cases. New unknown hazardous scenarios

may always arise each time changes are introduced in the system or in the deployed environment.
ISO 21448 [54] classifies the relevant scenarios of a use case into four areas (Fig. 3.8). The aim of this classification is to provide a conceptual abstraction that can define the overall goal of the SOTIF process that is reduction of the known/ unknown and hazardous scenarios. The aim of the activities carried out under SOTIF is to increase area 1 as much as possible (Fig. 3.8). Any use case on which SOTIF is applicable can consist of known and unknown scenarios. Through scenario discovery and identification of the use case, unknown and hazardous scenarios can be reduced. If relatively large area 2 and 3 are present, the existence of unreasonable risk can be argued. The goals of the SOTIF process with respect to Area 1, Area 2, Area 3 and Area 4 and relevant scenarios are:

- Area 1: To improve SOTIF, this area should be maximized. This refers to the evaluation of known and unknown hazardous scenarios.
- Area 2: To improve SOTIF, this area should be minimized to an acceptably small level. Moreover, by improving the functionality, the hazardous scenarios can be shifted to area 1. SOTIF refers to verification activities for this purpose.
- Area 3: To minimize this area, considerable effort is required to find unknown hazardous scenarios. SOTIF refers to validation activities for this purpose.
- Area 4: Even though area 4 is not hazardous, while performing minimization on the area 2 and 3, numerous scenarios from area 4 will be discovered and identified.

The unknown areas are related to the following category of scenarios.

1. The potential triggering conditions are defined but system response is unknown
2. Unknown triggering conditions
3. Combination of known triggering conditions resulting into potential unknown triggering conditions (e.g., combination of traffic and occlusion effects)

Random testing is recommended to uncover unknown hazardous scenarios from area 3 to known hazardous scenario of area 2 (Fig. 3.8). In the initial state, some potential functional insufficiencies and triggering conditions from the environment have been identified through safety analyses (clause 7). This identification corresponds to known hazardous scenario i.e., area 2 of Fig. 3.8. Other functional insufficiencies under triggering conditions are then identified through validation activities. It can be inferred that identification of functional insufficiencies and triggering conditions provides the basis of the evaluation of the known hazardous scenarios corresponding to an ODD and validation activities further probe the ODD to identify unknown hazardous scenarios (Fig. 3.7). Clause 7, 10 and 11 of the SOTIF activities (Fig. 3.7) can be summarized as processes to identify functional insufficiencies and triggering conditions as well as evaluate the scenarios. This forms the basis for the evaluation of known scenarios while probing the ODD for unknown hazardous scenarios.

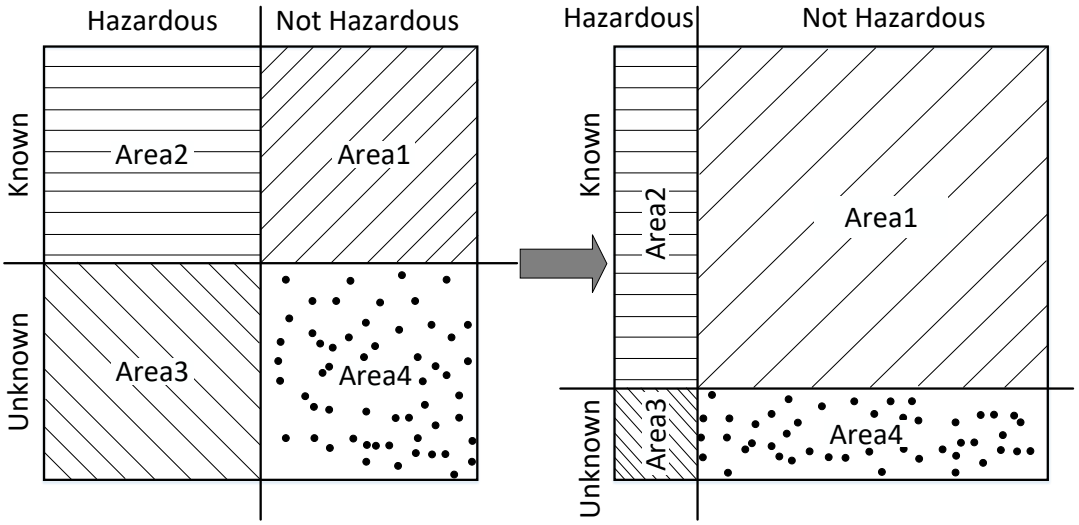


FIGURE 3.8: Evolution of the scenario categories resulting from the ISO 21448 activities [54]. Specifically, with the discovery of known and unknown hazardous conditions while also providing SOTIF improvement measures, the scenarios can be moved into the known and not hazardous (safe) area.

3.4.2.3 Measures to Improve SOTIF

Measures to improve SOTIF constitute clause 8 of the standard [54]. The purpose of this clause is to specify and apply SOTIF improvement measures. System refinement through incorporating SOTIF measures is performed, system specifications are updated accordingly and risks are evaluated. Possible SOTIF improvement measures considered are as follows.

1. System Modification
   (a) Improved sensor calibration and installation
   (b) Sensor blocking detection
   (c) Improved sensor technology
   (d) Sensor diversification
   (e) Improved recognition algorithm
   (f) Recognition of known unsupported environmental condition
2. Functional Restrictions
   (a) Restriction in steer assist torque
   (b) Limitation of the ODD
   (c) Restriction of the driving policy
   (d) Restriction during time of day (e.g., to avoid camera blindness during day and night)
3. Handing over of the authority
   (a) Dynamic driving task fallback strategy
   (b) HMI modification

3.4.3 SOTIF Analyses

Modelling the dependencies and influencing factors of the system to assess performance limitations and consequently the relevant uncertainties is important for SOTIF argumentation [54]. Such models can provide valuable insights on the functional performance of the system during development.

Safety analysis methods are adapted in ISO 21448 to identify and evaluate functional insufficiencies, triggering conditions and their dependencies. In the following sections, a critical review of CTA, SOTIF oriented FMEA and STPA is presented.

3.4.3.1 Cause Tree Analysis

The standard references a tailored version of FTA [129] for SOTIF implementation, called CTA. This analysis can determine root causes of the events, thus can be used for identification and understanding of the triggering conditions of a specific hazardous event. Traditional FTA [129] takes the following assumptions.

Assumption 1 *A cause and effect relation is deterministic in FTAs. The causality defined passes through AND gate [129].*

Assumption 2 *FTA is based on the Bernoulli process model i.e., a finite or infinite sequence of binary random variables, so it is a discrete-time stochastic process that takes only two values, canonically 0 and 1 [119].*

Assumption 3 *A basic event is considered as an independent event*[a] *[129].*

[a] Independent event is an event whose occurrence is not dependent on any other event.

Since ISO 21448 does not provide argumentations against these assumptions, naturally they can be translated to CTA.

Assessing the performance limitations under triggering conditions through CTA may pose some challenges, given the Assumptions 1, 2, 3.

For example, as presented in Fig. 3.5, illumination's effect on the FN probability of a perception sensor (e.g., LIDAR) is rather complex and confounded by weather. Moreover, *illumination=high* may affect the FN occurrence randomly i.e., it may or may not cause high FN probability in two different instances. Modelling conditional relations that also emanate inherent randomness using traditional FTAs is somewhat challenging (Assumption 1). Inhibit gate [129] provides means to model such conditional relation by including conditional probability. However, their usage to some extent is limited in literature.

As CTA is based on the Bernoulli process model (Assumptions 2), its ability to model continuous and multi-state events is restricted. Phenomena present in the open context are generally continuous in nature e.g., *illumination*. Modelling open context phenomena as CTA events thus require discretization. Discretization of variables may result in loss of information [35].

Basic events modelled in CTA are considered independent (Assumption 3). For many phenomena present in the open context independence cannot be guaranteed e.g., if weather and road surface conditions are taken as basic events, independence cannot be guaranteed (Fig. 3.2). Beta factors are used to model dependence between basic events [8].

FTA models the probabilistic occurrences of events using probabilities, allowing a representation of inherent randomness in occurrences. However, the causal effects in these FTA are still modelled through AND gate.

Fuzzy Fault Tree Analysis (FFTA) provides a conceptual framework to allocate lack of knowledge and vagueness in the probability values [73]. However, FFTA lacks in providing the overarching solutions to the complex causal relation modelling and iterative inclusion of novel causal factors in the analysis.

Open context and emergent nature of the environment and the HAD vehicles may require an iterative process with abilities to include novel triggering conditions and emergent properties identified during the system lifetime (till decommissioning). This becomes especially important when testing data becomes available for analysis. Generally, FTA is constructed as a one-time safety artefact. New evidences about the system and the context are not accommodated in the analysis. Moreover, FTA is generally considered to be an expert-based technique. Acquired testing datasets are seldom used to further improve the existing FTA.

SOTIF defines this *lack of knowledge* about the environment through *unknown unsafe scenarios* and advocates the discovery of these scenarios through V&V procedures. However, it lacks in providing iterative loops to safety analyses techniques to accommodate triggering conditions identified through evaluation of unknown unsafe scenarios.

3.4.3.2 SOTIF FMEA

The standard advocates an FMEA tailored to SOTIF analysis. A major shortcoming of FMEA is that only single point failure modes are evaluated. Multiple point failures cannot be studied through this method. In the example quoted in the previous chapter (Fig. 2.3), the failure of power supply can be considered as single point of failure for the system under consideration. For complex systems operating in the open context multiple point failures are equally important. Especially in the case perception system of HAD functions, multiple dependencies on triggering conditions can lead to frequent performance limitation of individual perception sensors.

FMEA does not provide a clear methodology on the causal relation of the prescribed causes of the failure modes. Moreover, the modelling of causal relations is limited in FMEA. The analysis only provides cause and effect semantics. It does not explain how failure modes, their identified causes and effects are related. This information can be helpful for SOTIF analysis as it assists in identification and evaluation of functional insufficiencies and triggering conditions [54].

3.4.3.3 SOTIF Oriented STPA

ISO 21448 also references STPA for SOTIF analysis. To this dissertation, the last step of STPA i.e., identification of loss scenarios (Sec. 2.2.3) is deemed the most relevant. SOTIF defines this step as identification of causal scenario that may lead to hazards and the corresponding causal factors (i.e., triggering conditions) [54]. The triggering conditions are determined by identification of functional insufficiencies through the analysis of the technical design. However, in the light of the challenges described for analysing the safety of the HAD vehicles (Ch. 3), some of the limitations presented in the Tab. 3.2 remain.

Causal Scenario	UCA (Hazardous Behaviour)	Functional Insufficiency	Causal Factor (Triggering Condition)
CS-1	UCA-1: Highway pilot does not provide a brake command when a forward collision is imminent.	FD-1: Highway pilot erroneously believes that there is no collision imminent due to inadequate feedback: Relative position, speed, acceleration, direction to an obstacle.	TC-1: Sensors mounted incorrectly, sensor focus or position compromised, sensor blocked, etc. TC-2: Feedback delayed and not received in time because the bus is busy, inadequate message priority or arbitration, EMI, etc.
...	...	...	...

TABLE 3.2: Identification of causal factors (triggering conditions) as defined in ISO 21448 [54].

Though causal factor identification is provided (Tab. 3.2), this relationship can be at best described by randomness, thus taking some of the functional insufficiencies and triggering conditions as random variables. STPA lacks the representation for random variables. Moreover, defining causality for random variables is not a deterministic notion. Thus, the conditional relation notion cannot be defined by a deterministic proposition.

Another shortcoming is the open context nature of the environment. This indicates the presence of multiple unknown triggering conditions for a specific system and context. Consideration of all the triggering conditions in (Tab. 3.2) based solely on the expert knowledge is challenging. Moreover, STPA does not contribute to the knowledge acquisition process of identification, modelling, quantification and validation of novel triggering conditions and performance limitations.

3.5 Summary

Non-linearity and emergent behaviour of the system results in randomness and variability between the functions of HAD vehicle while semi-permeable boundaries results in scenarios where a perfect system-environment description is challenging to formalize.

HAD vehicles rely on the sensing functions to interpret the open context. The open context introduces variabilities in the occurrences of the phenomena and their causal relations. Analyses of HAD functions also requires modelling of all the safety relevant aspects present in the open context. Complexity of the HAD vehicles and how they interact with the open context results in the influencing factors with complex causal relations. Modelling all the influencing factors is challenging, introducing a lack of knowledge about system, open context and their underlying interactions. In order to determine these causal relations, more in-depth knowledge on the

process is required. Traditional safety analysis techniques do not provide the agility required for the iterative processes.

FTA, FMEA and STPA are safety analysis techniques that rely on expert knowledge for modelling and analysing hazardous behaviour, undesired events and their causal factors traditionally [129, 117, 70]. Some extensions to these techniques exist quantification of probabilities through data e.g., to represent the individual events and joint probabilities of events [129] or to allocate lack of knowledge and vagueness in the probability values [73]. The assumptions taken by these analysis methods are seldom held e.g., FTA assumptions are weak for complex systems and open context analysis. They also do not mandate a modular and iterative modelling scheme that can assist in modelling and guided discovery of novel triggering conditions through expert and data engineering-oriented techniques.

In order to assess the HAD vehicles safety, a vast number of scenarios comprising of triggering conditions need be to analysed owing to the open context. Moreover, the system behaviour and complex causal relation further convolute the modelling and analysis problem.

Experts can only provide models based on their limited knowledge, while data generated through real world experiments is marginal, joint and observational in nature. Confounding phenomenon, posed by the unobserved latent variables may invalidate the analysis.

ISO 21448 [54] provides structured guidelines for automated driving functions safety by identification of triggering conditions through analyses, V&V of known and unknown unsafe scenarios. However, solutions on the provision of causal effects of triggering conditions on the performance limitation are not provided by SOTIF. For example, rain and snow are known to affect RADAR performance. In order to measure this performance limitation, a performance metric needs to be selected and a corresponding causal effect should be measured and quantified. This amounts to complex interactions challenges discussed earlier (Sec. 3.3.2). ISO 21448 [54] provides a list of such dependencies in terms of scenario factors but does not provide concrete steps to model these scenario factors.

Moreover, evaluation of HAD vehicle requires characterization of many influencing factors owing to the open context. This results in maximizing area 1 (Fig. 3.8). This evaluation requires discovery and identification of all the dependencies, triggering conditions, functional insufficiencies and performance limitations that can lead to hazardous behaviour. SOTIF provides methodologies to cover the search space of the influencing factors, including multiple analyses techniques, however, does not provide any argumentation on the exhaustiveness of the used techniques. This argument leads to completeness concerns when functional insufficiencies are evaluated using traditional safety analysis techniques. Incompleteness in the safety evaluation of functional insufficiencies may originate from challenges of HAD vehicle safety. Challenges discussed above result in uncertain models.

The challenges discussed in the previous section can be summarized as follows.

Proposition 1 *There exists randomness in models of the causal relations of **functional insufficiencies** owing to the attributes of complex system, open context and their interactions. Thus, deterministic solution provision for analysis of triggering conditions and functional insufficiency may become infeasible.*

Proposition 2 *There exists randomness in modelling the occurrences of* ***triggering conditions*** *and* ***functional insufficiencies*** *owing to the attributes of complex system, open context and their interactions. Thus, deterministic solution provision for analysis of triggering conditions and functional insufficiency may become infeasible.*

Proposition 3 *There exists lack of knowledge in the existence of triggering conditions pertaining to a given functional insufficiency owing to the open context nature of the environment.*

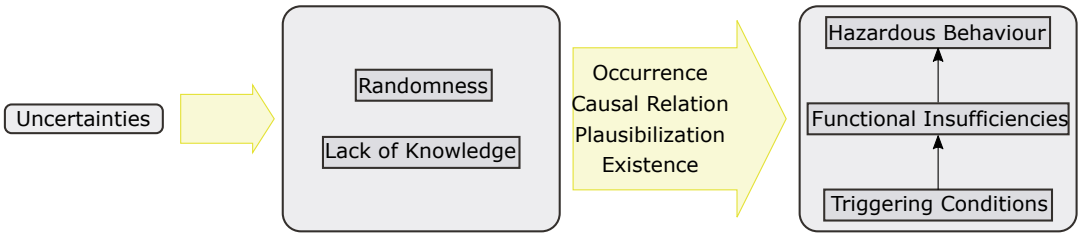


FIGURE 3.9: A high level definition of uncertainties involved in modelling the triggering conditions and functional insufficiencies.

Randomness, variability and lack of knowledge are terms associated with uncertainty [49]. As a generic concept of the unknown, incomplete or imperfect knowledge, it has been classified in various contexts in the literature [131, 24, 30, 41]. In this thesis, a categorization is provided based on the notion of randomness and imperfect knowledge. Current safety analyses techniques also mandated by SOTIF do not provide scope to model and represent uncertainties.

4

Uncertainty Models for Modelling Safety of the Intended Functionality

"Not only does God play dice, but... he sometimes throws them where they cannot be seen."

– Stephen Hawking, *English Theoretical Physicist and Cosmologist*

This chapter provides an uncertainty categorization deemed suitable for HAD systems. It also discusses and provides safety analyses models to represent uncertainties. In Sec. 4.1, chapter contributions are summarized. An overview of uncertainties is provided in Sec. 4.2. In Sec. 4.3, different system models and their ability to model uncertainties are discussed. It is then followed by graphical models for uncertainties (Sec. 4.4). In Sec. 4.5 a detailed understanding of the semantics of different types of uncertainties is presented. Sec. 4.6 provides how SOTIF can be modelled and analysed in the various graphical modelling frameworks. A critical review is also given for the discussed modelling frameworks. In the last section, a summary of this chapter is provided (Sec. 4.7).

4.1 Chapter Contribution

In this Ch., the following contributions are made.

(S1) Categorization of uncertainties to address SOTIF analysis

(S2) Provision of safety analyses models to represent uncertainties

(S3) A representative SOTIF example modelled in the proposed models to represent uncertainties

Some of the sections that constitute this chapter has been published as scientific contributions [41, 3].

4.2 Uncertainty Categorization

Uncertainty refers to the concept of insufficient or unknown piece of knowledge [49]. In the literature, it is treated as the generic notion of imperfect, incomplete and unknown knowledge. Moreover, it is categorized as *aleatory* and *epistemic* [131, 30]. In this dissertation, uncertainty is considered as a notion in representing knowledge through models. Depending upon the origins of uncertainties in the models to analyse HAD functions safety, a categorization between aleatory, epistemic and ontological uncertainty is made and discussed in the following.

4.2.1 Aleatory Uncertainty

Definition 7 *Aleatory uncertainty can be regarded as randomness of a process represented by a system model.*

Aleatory uncertainty is considered to be irreducible for a given choice of a probabilistic model and is quantified by probability distributions [30].

Let us consider a perception system consisting of a LIDAR and deep neural network that detects objects on highways. It is also considered that the study is conducted to assess the FN probability of the perception chain under different factors as shown in the initial example (Fig. 3.5). The occurrence of these variables can be represented by a random variable [72]. If data is gathered and labelled, the relative frequency of occurrence of the variables involved in the measurement of world model can be represented using probability distribution e.g., *P(FN)* represents the probability distribution of FN of LIDAR. This probability distribution represents the aleatory uncertainty of the world model.

4.2.2 Epistemic Uncertainty

Definition 8 *Epistemic uncertainty is associated to the lack of knowledge about the system model and the inexact encoding of physical systems to models.*

With epistemic uncertainty, the lack of knowledge can be represented. Taleb [121] refers it to as the known-unknown of the model. Epistemic uncertainty has also been characterized as conditional entropy [106, 51], i.e., the difference of information between modelled and physical system. A model is the abstraction of reality [126]. Since epistemic uncertainty is defined as the general lack of knowledge about reality, a unique and distinct measurement and representation is not available in the literature.

Epistemic uncertainty can be represented as the approximation of real and unknown probability distribution through the parameter taken in the example presented in Fig. 3.5. It can also be represented as the missing element in the model. Summarizing, the objective view on epistemic uncertainty stems from the problem definition and its proposed solutions.

4.2.3 Ontological Uncertainty

Definition 9 *Ontological uncertainty can be defined as a condition of complete ignorance in the model of a relevant aspect of the system.*

This has also been termed as the unknown-unknown [121], the state of we do not know that we do not know. Ontological uncertainty is based on the study of existence. It can be inferred as the lack of knowledge about the existence of relevant aspects in our model representation.

Ontological uncertainty can be represented as novel causal factors, never observed before. It can also represent the unknown states of different causal factors. For example, Volvo self-driving cars were unable to detect kangaroos in Australia [137]. Volvo reported that their "large animal detection system" was unable to detect kangaroos owing to their irregular method of movement. It can be termed as the *"Black Swan Event"* of the Volvo perception detection system.

Consideration of ontological uncertainty as a separate artefact is valuable for safety analysis of HAD vehicles, as it requires different means of representation and mitigation [41]. In the case of HAD vehicles, ontological uncertainty can never be fully eliminated during the vehicle lifetime, owing to their deployment in the open context. For example, e-scooters were not part of road traffic a decade ago. Even though epistemic and ontological uncertainties originate from lack of knowledge, a general distinction can be made between model parameters (epistemic) and model correctness (ontological) to segregate the two uncertainties. In order to represent the ontological uncertainty in the system model, the notion of unknown state was introduced [41].

4.3 System Models and Uncertainty Representation

Models are the abstract representation of reality [116, 126]. Rosen provides a formal basis for models [96]. He argues that modelling is an isomorphic encoding $\epsilon_{A,B}$ and decoding $\delta_{A,B}$ of relevant properties of the natural system (reality) into formal systems. The causality in the natural system is thereby mapped to logic inferences in the model. The formal systems can be attributed to mathematical equations, probability distributions or, in the context of safety engineering, failure analysis models (e.g., FTA).

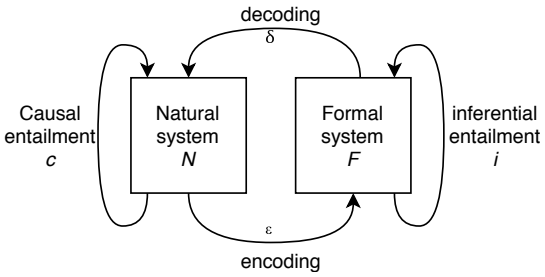


FIGURE 4.1: Modelling relation between a natural system N and formal system S. Inferential entailment i represent the causal entailment c if the encoding ϵ and decoding δ is isomorphic consisting of two ideal point masses (planet 1 and 2) and two formal systems as models [96].

Multiple encoding principles can be used to encode natural to formal systems, i.e., there are multiple models available for a system. They serve the needs of the modeller, industry and the use case they are intended for. Since this thesis intends to model the uncertainties as discussed in the previous section, a distinction between deterministic and probabilistic[1] models is made.

From the deterministic model, a distinct outcome can be inferred for a given set of input parameters. While the probabilistic models infer statements about probabilistic outcomes for a given set of inputs. Deterministic models dominate representation and causality for natural system in general. However, to model uncertainties, probabilistic models are a more feasible choice.

In the example of the previous section (Fig. 3.5), the reality is the *HAD vehicle operating in an open context*. The emphasis is on the HAD vehicle's functions performance under the influence of various causal factors e.g., effect of rain on the LIDAR

[1]The term "probabilistic" is a term used only for probability theory. However, in this chapter it is also used to cover Belief Theory [104].

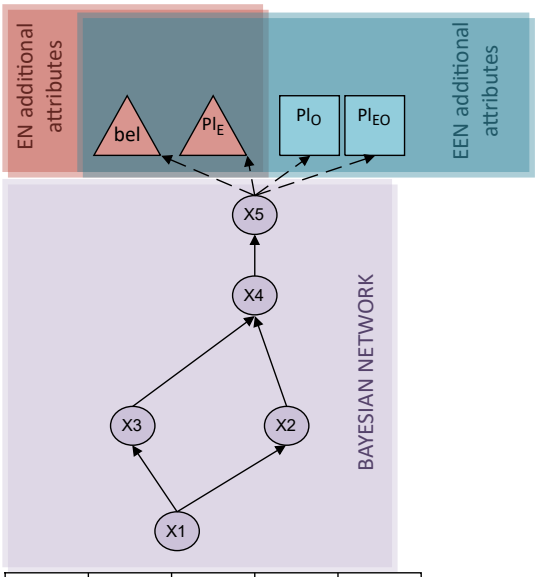
performance. The behaviour of the system can be described by various mathematical equations [45, 50]. These equations are the representative of a deterministic model that infers the causality of LIDAR performance under rain. For different initial conditions and parameters, various states of LIDAR performance can be calculated. The calculation of future state of reality is possible with a hypothetical Laplace demon [69]. Laplace defines it as an entity, which has the perfect knowledge of present states and can perfectly predict the future states. In terms of Rosen's formal basis of models, this corresponds to formal system with perfect encoding $\epsilon_{A,B}$ and causal mechanism of the physical system. In reality, however models are not perfect representation because of the various practical and theoretical reasons [78, 33].

Another possibility to define the effect of rain on the LIDAR performance is by a probabilistic model using either frequentist [22] or Bayesian [84] approach. This means collecting observational/experimental databases on rain and LIDAR performances and using it to erect a probabilistic model. With an infinite amount of an observational database or a randomized controlled experiment for experimental database, the exact probability functions can be produced.

Deterministic and probabilistic models fulfil the modelling relations and enable the modeller to draw meaningful conclusions about the system. Selection between deterministic and probabilistic model is based on the need of the analysis as well as availability of databases and functions. Probabilistic models become an inherent choice for systems with limitations of theoretical determinism. In applications like HAD functions, the challenges discussed to model the HAD vehicles' safety (Sec. 3.3), necessitates the use of probabilistic models.

4.4 Probabilistic Graphical Models

Probabilistic graphical models (PGMs) [66] are a graphical representation that encode a joint distribution over a high-dimensional space. PGMs consists of nodes and edges (Fig. 4.2). Nodes correspond to the variables and edges correspond to direct probabilistic interaction between them as shown in Fig. 4.2.



X4	X5=Fail	X5= Success	X5=Fail or Success	X5= Unknown
Fail	0.7	0.1	0.1	0.1
Success	0.5	0.2	0.1	0.2
Fail or Success	0.2	0.3	0.4	0.1
Unknown	0.05	0.05	0.3	0.6

FIGURE 4.2: Example of Bayesian Network (BN): A BN with nodes (X1,...,X5) and edges represented by arrows. Evidential Network (EN) adds two additional nodes of belief and plausibility. Extended Evidential Network (EEN) adds further nodes to represent notion of ontological uncertainty. Exemplary conditional probability table for X5 given.

The graphical structure provides information on two different aspects. The first aspect provides a set of independence that hold in the distributions; it takes the form of **X** is independent of **Y** given **Z**, denoted **(X⊥ Y | Z)** for some subsets of variables **X,Y, Z**. For example, in the Fig. 4.2, **(X5⊥ X1,X2,X3 | X4)**.

The second aspect provides the factorization of joint probability distribution in order to compactly represent a high-dimensional distribution. The overall joint probability distribution can be defined as a product of these factors.

4.4.1 Representation, Inference, Learning

PGMs utilize the graphical structure to emphasize the fact that variables tend to interact directly only with a subset of other variables. This modelling framework has many advantages. It allows representation of astronomically large distribution with manageable factors. Such factorized representation is transparent, in a way that a human can evaluate and understand the underlying semantics of joint probability distributions.

The graphical structure also allows answering the queries through inference, stipulating posterior probabilities of the variables given evidence on other variables for a given mathematical query. For example, if rainy weather, low illumination and wet road scenario is observed, reflection may become the variable of interest as it can cause higher FN.

PGMs enable the construction of the model either by the human expert or automatically by learning techniques through data[2], thus providing an approximation over the experiences. In the following four PGMs i.e., BN, CBN, EN and EEN are discussed. EN and EEN discussed in this thesis are an extension to the BN/ CBN.

4.4.2 Bayesian Networks

Bayesian Networks (BNs) [88] are frequently used tools in the dependability research [132], [21].

Definition 10 *A BN is a Directed Acyclic Graph (DAG) that comprises nodes and edges. A node represents a random variable $(X_1, \ldots, X_n)$, while the edges run from the parent node (pa) towards the child node (ch). This combination of nodes and edges represents the structure of BN. The dependencies between two nodes are modelled using conditional probability distributions $P(ch \mid pa)$ [66].*

If the *Markovian condition* is satisfied, the BN can be written as follows [66].

$$P(X_1, \ldots, X_n) = \prod_{i}^{n} P(X_i \mid pa(X_i)) \tag{4.1}$$

Where $P(.|.)$ is known as conditional probability and is defined by Bayes rule as follows.

Definition 11 *Given two variables X and Y in a DAG (G), the association[a] of X on Y is given by.*

$$P(y|x) = \frac{P(x|y) * P(y)}{P(x)} \tag{4.2}$$

[a]The term association is used for conditional probability in light of Judea Pearl's causational ladder [87].

BN is effective in modelling uncertainty and probability reasoning of a system. More specifically, it models the aleatory uncertainty of the system model [41]. BN can be constructed by defining a DAG along with joint probability distributions governed by Eq. 4.1. Both the DAG and probability distributions can be either provided by experts or learned through data (Fig. 4.3). In this thesis, the initial structure (DAG) is provided by the human expert while the joint probability distributions are learned by data, unless stated otherwise.

BN exploits the dependence relationship through the local conditions in the model to perform uncertainty analysis for prediction and classification of influencing factors. It also assumes two completeness conditions [124].

Assumption 4 *The probability distributions are known to have an acceptable level of accuracy and precision.*

Assumption 5 *The random variables are independent or the dependence is known and modelled in the BN.*

[2]Throughout this dissertation, the initial structure (DAG) is provided by the human expert. The joint probability distribution is learned through data, unless otherwise stated.

It has been argued that both conditions (Assumption 4 and Assumption 5) are rarely fulfilled [124].

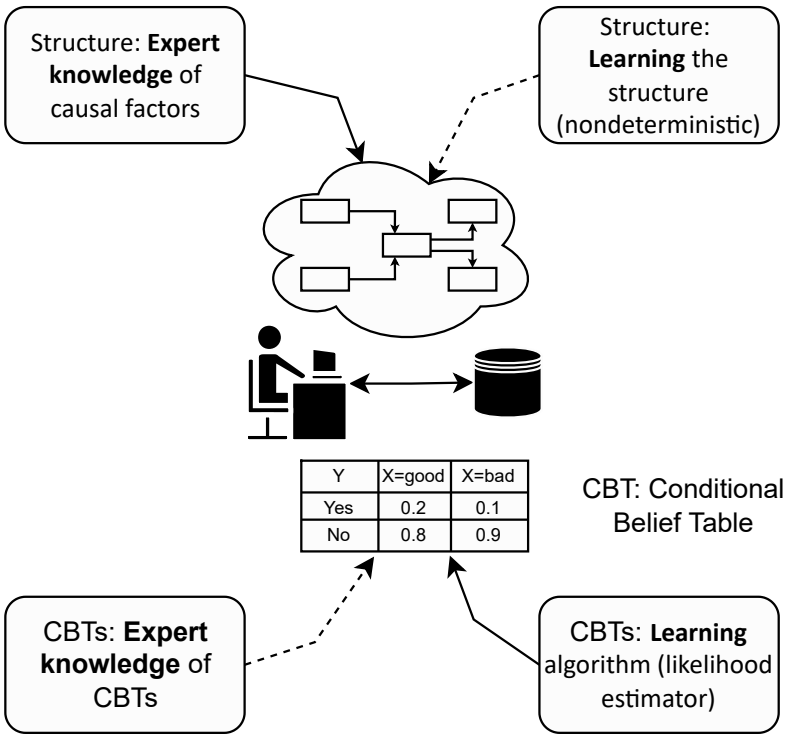


FIGURE 4.3: BN construction: The BN structure (DAG) along with joint probability distributions can be elicited by human expert or learnt through data. In this work, the initial structure is provided by human experts while the joint probability distributions are learnt by data, unless stated otherwise.

4.4.2.1 Causal Bayesian Network

The independence assumptions in BNs does not necessarily imply causation. Traditional concepts of probability theory only infer associations but they lack causal relationships. They are valid for any set of variables. However, the prevalence of BNs in various field stems from the causal interpretation [87]. Pearl augments these concepts by introducing a causal inference framework [87, 89]. He provides the representation of causal relationship with graphical models and Bayesian statistics. In this section, an overview of the concepts of causal theory is provided.

Pearl argues that the causal relationship can be represented by causal structures. A causal structure is a DAG represented by a set of variables. A distinction is made between *endogenous variables*, which are determined by other variables in DAG and *exogenous variables* which define errors or disturbances. Exogenous variables are considered out of scope of this thesis. An edge in the DAG remarks as a causal influence. Conditional independence is considered as the primary source of expressing knowledge about the world [87, 88]; Pearl considers this conditional independence the by-product of the causal relationships [87].

A *causal model* defines how each variable is influenced by its parents. More precisely, the model (a DAG) consists of joint probability distribution $P(X_1, \ldots, X_n)$ which is a function of parents $pa(X_i)$ for each variable X_i.

Pearl introduces a mathematical operator called *do-operator* to analyse the causal effect. The do-operator simulates a physical intervention. For a set of variables

X, this is implemented by deleting all the functions in the model that define the variable in X and assigning $X = x$ in the other functions(Fig. 4.4). Therefore, post intervention distribution of an event Y can be described by the following relation.

$$P(Y = y|do(X = x)) = P(Y = y|do(x)) \tag{4.3}$$

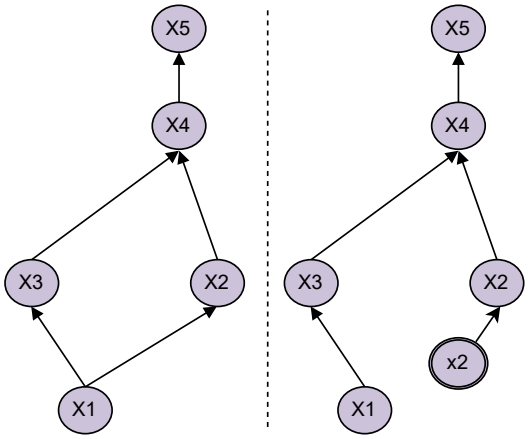


FIGURE 4.4: Example of a Causal Bayesian Network (CBN) pre- and post-intervention at variable $X2$. All the arrows coming in to $X2$ are sliced and the variable is set as $X2 = x2$.

The quantity (Eq. 4.3) is defined as the causal effect of x on y. In the real world, intervention is performed through experimental setup e.g., randomized controlled experiments [15]. However, in general, randomized controlled experiments are not feasible in the field of automated driving for the following reasons.

1. Ethical: Randomized experiments cannot be conducted for accident data.
2. Infeasibility: Randomizing the perception system on different OEM vehicles.
3. Impossibility: In some cases, the experiment cannot be altered.

If Markov conditions are fulfilled, the causal effect of X on Y is identifiable, if X, Y and parents of X, $PA(X)$ are measurable [89]. Eq. 4.3 then can be rewritten as follows.

$$P(Y = y|do(X = x)) = \sum_{pa(x)} P(Y = y|X = x, PA(X) = pa(x))(PA(X) = pa(x)) \tag{4.4}$$

However, there may exist a different set of variables apart of the parent nodes, that are sufficient to measure the causal effect by fulfilling the so called *back door criterion*. The back door criterion states that a set of variables Z fulfil the adjustments required to calculate a causal effect, if Z does not contain any descendants of X and variables in Z blocks all the paths from X to Y that contains an edge into X [87, 89]. Formally, this can be written as follows.

Definition 12 *$Causal\ Effect_{Backdoor\ Criterion}$: Given two variables X and Y in a DAG (G) with a set of variable Z which satisfy backdoor criterion, the causal effect of X on Y*

is given by.

$$P(y|do(x)) = \sum_{z} P(y|x,z)P(z) \tag{4.5}$$

In the language of causal theory, the set of variables Z achieves *d-separation* between X and Y. If the set of variables Z is an empty set, Eq. 4.5 and Eq. 4.2 become equivalent. Since causal structures are considered as DAGs the concept of causal theory can be directly translated to BN, resulting in Causal Bayesian Network (CBN).

There are several advantages of CBN over purely associational model. CBN are based on more meaningful, reliable and accessible judgments. For example, construction of two variables *weather* and *FN* as a CBN is more meaningful than two random variables. CBN also assists in the justification of independence i.e., the independence relation *(X5⊥ X1,X2 | X3,X4)* defined in Fig. 4.2 can be easily converted to a more meaningful one involving causal relationships; that the influence of *weather* on *illumination* is mediated by *reflection* and type of *road*. Another advantage of CBN is the ability to represent an inferential change. Any reorganization among the causal mechanism can be translated to the CBN and inferred. The reorganization of causal mechanism rest on the assumption that any change in the parent child relationship is modular i.e., it does not impact the relationship of other variables. The reorganization in the CBN allows the modeller to model and infer intervention queries with minimal information. CBN are much more informative than their "associational" counterparts (BN). BN only provides joint distribution which in turn tells us the probabilities of events and change in the probabilities upon observations. On the other hand, CBN also provides change in probabilities upon external intervention. The topic of intervention and its importance for safety of the HAD vehicle and deducing design principles on it will be discussed in detail in the subsequent chapter.

Probability theory and BN are considered sufficient to represent aleatory uncertainty [111]. However, sufficiency of probability theory to represent epistemic uncertainty has been challenged by some authors [36]. In the next section, an introduction to Evidence Theory [104, 28] is provided followed by a formal definition of the DAGs based on the theory.

4.4.3 Evidential Networks

The Dempster and Shafer Theory (DST) or Evidence Theory (ET) is a mathematical theory that structures phenomenon by degree of beliefs (belief masses) on events or states [28, 104]. Conceptually, DST can be viewed as a generalized Bayesian Model [113]. This characteristic increases its applicability on the safety analyses, where BN algorithms are used [110]. DST comprises the following three attributes.

4.4.3.1 Frame of Discernment

Consider the multi-state analysis outcome with n mutually exclusive and exhaustive states. The frame of discernment Ω is the finite set of such elements as follows.

$$\Omega = \{y_1, y_2, \ldots, y_n\} \tag{4.6}$$

In DST, the Basic Belief Assignment (BBA) is calculated on the power set of frame of discernment.

$$\begin{aligned} 2^{\Omega} = \{&\emptyset, \{y_1\}, \{y_2\}, \ldots, \{y_n\}, \ldots, \{y_1, y_2\}, \\ &\ldots, \{y_1, \ldots, y_n\}\} \end{aligned} \tag{4.7}$$

4.4.3.2 Basic Belief Assignment

Information on the outcome states (power set) is assigned by belief $m(A)$ with the following properties $m : 2^{\Omega} \rightarrow [0, 1]$ and

$$m(\emptyset) = 0 \tag{4.8}$$

$$\sum_{A \in 2^{\Omega}} m(A) = 1 \tag{4.9}$$

where A is the subset of the power set of frame of discernment. BBA can be seen as an alternative to probabilities. In this publication the term BBA and belief mass for DST, EN and EEN and probabilities for BN or CBN parameters is used. The subsets fulfilling $\{A \in 2^{\Omega} : m(A) > 0\}$ are called focal elements. Full knowledge can be represented by assigning masses to singleton sets of Eq. 4.7, while assigning mass $m(\Omega) = 1$ represents total ignorance [5]. Eq. 4.8 constrains the outcome elements to the closed world assumption [93].

4.4.3.3 Belief and Plausibility Measures

These measures provide upper and lower bounds on the BBA in DST with the following mathematical structures.

$$bel(B) = \sum_{A|A \subseteq B} m(A) \tag{4.10}$$

$$pl(B) = \sum_{A|A \cap B \neq} m(A) \tag{4.11}$$

where B is the subset of the power set of frame of discernment. The difference between plausibility and belief function provides a notion of epistemic uncertainty [4, 109, 92]. The belief measure $bel(B)$ can be seen as sum of BBA of all the subsets of Ω that are *fully* in agreement with B, while $pl(B)$ can be regarded as sum of BBA of all the subsets of Ω that are *fully or partially* in agreement with B [5]. For singleton subsets of frame of discernment Ω, where BBA and belief functions are same, plausibility functions can model the lack of knowledge postulation. However, when categorized into ontological and epistemic, it becomes challenging to comprehend which uncertainty among epistemic and ontological is represented by the difference of unique plausibility and belief function.

Definition 13 *Evidential Networks are also DAGs which represent uncertainties as randomness (aleatory) and lack of knowledge (epistemic) [110]. They use nodes to represent random variables, edges to define direct dependence between nodes and conditional belief mass to quantify dependency. When a node is a root, a priori belief mass table is defined. Moreover, distinction is made for leaf node by providing belief and plausibility measures (Fig. 4.2). The dashed arrows signify the fact that there is no influence involved in those connections.*

4.4.4 Extended Evidential Networks

In this section, an approach that extends the representation of uncertainties using DST by incorporating ontological uncertainty is proposed. EN are extended to incorporate both epistemic and ontological uncertainties separately through EEN.
In this regard, the DST attributes are redefined as follows.

4.4.4.1 Frame of Discernment

Consider the multi-state analysis outcome with the inclusion of ontological uncertainty through state u [41]. The state u refers to all those states that may not have been considered during system design/ analysis. Eq. 4.6 can be rewritten as

$$\Omega = \{y_1, y_2, \ldots, y_n, u\} \tag{4.12}$$

In DST, the BBA is performed on the power set of frame of discernment.

$$\begin{aligned} 2^{\Omega} = \{\emptyset, \{y_1\}, \{y_2\}, \ldots, \{y_n\}, \{u\}, \ldots, \\ \{y_1, y_2\}, \ldots, \{y_1, \ldots, y_n, u\}\} \end{aligned} \tag{4.13}$$

Further, three subsets of Eq. 4.13 are also defined as follows.

$$\begin{aligned} E = \{\{y_1, y_2\}, \{y_1, y_3\}, \ldots \{y_1, y_n\}, \ldots, \\ \{y_2, y_3\}, \ldots \{y_1, \ldots, y_n\}\} \end{aligned} \tag{4.14}$$

$$O = \{\{u\}\} \tag{4.15}$$

$$\begin{aligned} EO = \{\{y_1, u\}, \{y_2, u\}, \ldots \{y_1, y_2, u\}, \ldots \\ \{y_1, \ldots, y_n, u\}\} \end{aligned} \tag{4.16}$$

Here Eq. 4.14, 4.15 and 4.16 represent the epistemic, ontological as well as mixed epistemic and ontological uncertainty sets, respectively.

4.4.4.2 Belief and Plausibility Measures

As it is discussed in the previous section, belief measure $bel(B)$ can be viewed as sum of BBA of all the subsets of Ω that are fully in agreement with B and do not contribute to uncertainties, hence belief measure bel (Eq. 4.11) remains the same. The following presumptions about the Eq. 4.13 before defining plausibility functions are taken.

1. All singleton subsets are considered exempted from the uncertainty except u.
2. Element u is considered as ontological uncertainty (O).
3. Non-singular subsets not containing u are considered epistemic uncertainty of the system model (E).
4. Non-singular subsets containing u are considered mixed ontological and epistemic uncertainty of the system (EO).

Based on the above presumptions and Eq. 4.14-4.16, the multiple plausibility functions to individually characterize uncertainties in the analysis outcome are defined. $\{\forall B : B \subset 2^{\Omega} \wedge |B| = 1\}$

$$bel(B) = \sum_{A|A \subseteq B} m(A) \tag{4.17}$$

$$pl_E(B) = bel(B) + \sum_{\substack{A|A \cap B \neq \\ \wedge \quad A \in E}} m(A) \tag{4.18}$$

$$pl_O(B) = bel(B) + \sum_{\substack{A|A \cap B \neq \\ \wedge \quad A \in O}} m(A) \tag{4.19}$$

$$pl_{EO}(B) = bel(B) + \sum_{\substack{A|A\cap B\neq \\ \wedge\ A\in EO}} m(A) \tag{4.20}$$

The method that is presented here is applicable on the quantification of plausibility functions of the original frame of discernment "Ω" states only. Separate representation of epistemic and ontological uncertainty in EEN can assist in choosing the right improvement measure. For example, model refinement (changing model parameters) and model rediscovery (changing the model altogether) can be associated to epistemic and ontological uncertainty, respectively. Mixed epistemic and ontological uncertainty may serve the case where both model refinement and rediscovery require improvement. In other words, this categorization in the safety analysis may assist in the improvement measures by indicating the aspect to be improved (e.g., better parametrization of a model or redesigning a model all together).
Having provided with the approach to distinguish between epistemic and ontological uncertainties, the definition of EEN is provided as follows.

Definition 14 *Extended Evidential Networks (EENs) are DAGs. They represent uncertainties such as randomness (aleatory), lack of knowledge (epistemic) and state of complete ignorance (ontological). They use nodes to represent random variables, edges to define direct dependence between nodes and conditional belief mass to quantify dependency. When a node is a root, a priori belief mass table is defined. The leaf node represents the query of the network. Moreover, leaf nodes are distinct as belief and multiple plausibility measures are provided (Fig. 4.2).*

4.5 Semantics of Uncertainty Measurements

Uncertainty, if measured through a metric or measuring process, holds distinct semantics. A clear understanding of these semantics is necessary as it may also influence the SOTIF improvement measure. Semantics related to aleatory, epistemic and ontological uncertainties are discussed as following.

4.5.1 Semantic of Aleatory Uncertainty

The semantics of aleatory uncertainty are straightforward. Aleatory uncertainty measures the randomness of occurrence of an event. Such an event can be modelled as a random variable. A probability distribution $P(.)$ or conditional probability distribution $p(.|.)$ provides a representation for aleatory uncertainty. Aleatory uncertainty provides a representation of randomness and variability. BNs and CBNs are thus deemed sufficient for aleatory uncertainty representation.

4.5.2 Semantic of Epistemic Uncertainty

The underlying semantic of epistemic uncertainty is vague, owing to its definition of "lack of knowledge". At this stage, two major distinctions in the epistemic uncertainty elicitation process can be made: (1) Based on the subjective opinion of the expert. (2) Based on the measurement driven by mathematical principles.

The elicitation of epistemic uncertainty based on the expert opinion is the subjective estimation of the expert about the system models' representation of reality. It may represent the overall lack in the knowledge present in the system model.

A mathematical principle used to elicit epistemic uncertainty defines its objective view. What constitutes lack of knowledge is dependent on the mathematical

principle used. DST and EN/ EEN (Sec. 4.4.3) provide a well-equipped modelling tool to represent lack of knowledge, ignorance and vagueness of the system model. However, this tool does not answer the important query of the semantic behind that vagueness. Based on the underlying mathematical principle, multiple views in this regard have been taken. They include statistical variations [90], confidence interval [88], contingency sets [88] and entropy measurements [102]. However, measurement of epistemic uncertainty in this manner yields a distinct meaning to the "lack of knowledge" based on the mathematical principle.

4.5.3 Semantic of Ontological Uncertainty

The semantic of ontological uncertainty is also straight forward. It defines sheer ignorance about the system model and is based on expert guesses. In data learning approaches, the measures to represent ontological uncertainty can be appended later in the results. This is because the ontological uncertainty is considered to be not quantifiable through algorithms, as such quantification through algorithm makes it epistemic by definition.

4.6 SOTIF Analysis In PGMs

Performing SOTIF analysis in the PGMs requires a sound understanding of the nominal functionality of the system including the operational design domain [3]. Calculation of EN is considered out of the scope of this dissertation as EEN covers all three uncertainties. Specific to this dissertation, EEN is defined as an extension to CBN. In this regard, modelling a CBN is necessary to model EEN.

4.6.1 Modelling Steps

In this section, steps to perform a SOTIF analysis in PGMs are provided. Modelling steps are revisited in detail in the next chapters.

1. Performance Limitation Selection: Performance limitation related to SOTIF are selected. This may include but are not limited to the following.

   (a) The inability of the function to correctly comprehend the situation and operate safely; this also includes functions that use machine learning algorithms.

   (b) Insufficient robustness of the function, system, or algorithm with respect to sensor input variations, heuristics used for fusion, or diverse environmental conditions.

   Examples include False Positive (FP), FN and position trueness.

2. Triggering conditions Propositions: Conditions at scenario level that may lead to performance limitations are listed. This also includes environmental effects and foreseeable misuse [54]. Examples of triggering conditions may include the following.

   (a) Road/ traffic conditions

   (b) Weather conditions

3. CBN Construction: A CBN model is constructed out of the information from previous steps. The steps further taken are as follows.

(a) Hierarchical dependencies between hazardous behaviour, triggering conditions and insufficiencies of specifications and performance limitations are established.

(b) A CBN is constructed i.e., nodes representing hazardous behaviour, triggering conditions, insufficiencies of specifications and performance limitations while edges representing the dependencies.

(c) Belief masses for root nodes and Conditional Belief Tables (CBT) for intermediate nodes are assigned by a human expert to quantify the extent of dependencies.

(d) Belief masses of the leaf node are calculated by propagation.

4. EEN Extension: bel, pl_E, pl_O and pl_{EO} functions are calculated (Eqs. 4.17, 4.18, 4.20 and 4.19) for the leaf node states, thus constructing the EEN. Higher $pl_x \forall x \in \{E, O, EO\}$ and bel correspond to higher uncertainty.

5. SOTIF Improvement Measure: ISO 21448 requires improvement measures to address SOTIF. This includes avoidance and mitigation measures. While avoidance represents elimination of risk, mitigation measures consider reducing risk where avoidance is not possible.

4.6.2 Modelled Scene Description

Consider a HAD vehicle which is equipped with LIDARs for perception of the operational environment for classification of road actors. The perception function is designed to detect the cars only. LIDAR performance is chosen as the performance limitation measure as advocated by the SOTIF standard [54]. For this perception function, the experts provide the following propositions.

1. **Occlusion** in detection may influence LIDAR performance.

2. **Reflections** in the scene can affect the LIDAR performance.

3. **LIDAR performance** directly affects the car detection.

The experts believe that for low LIDAR performance we may have higher epistemic uncertainty about detection. This can be modelled using EEN. The reason for the epistemic uncertainty at low performance can be related to the available knowledge at this state.

4.6.3 Implementation

Fig. 4.5 shows the CBN resulting from the previous propositions. The CBN is extended by *belief* and *plausibility* nodes to also represent EEN concepts. Based on the constructed CBN, the extension of EEN with plausibility measures is shown in Tab. 4.1. The conditional belief tables are based on the expert opinion.

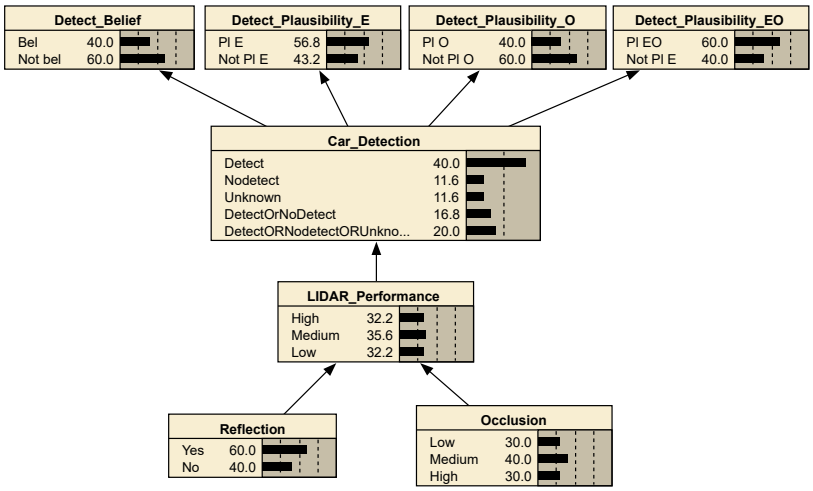


FIGURE 4.5: SOTIF analysis modelled in Causal Bayesian Network (CBN) with extension for Extended Evidential Network (EEN). The model provides the advantage of representation for aleatory, epistemic and ontological uncertainty. Probability can be propagated through different nodes. The difference between $pl_x \forall x \in \{E, O, EO\}$ and *bel* represent the different types of uncertainties as shown in Tab. 4.1.

4.6.4 Analysis and Observation

In the case of CBN, the only metric at the detection node level is the belief function value ($bel(A)$) i.e., the aleatory uncertainty measure. The conditional probability and related terms are out of the scope of this discussion. SOTIF improvement measures can be defined for the belief function e.g., for the low detection rate, a better sensor as replacement or further DNN training can be seen as potential SOTIF improvement measures.

Perception $A \subseteq 2^{\Omega}$	belief $bel(A)$	plausibility $pl_E(A)$	plausibility $pl_O(A)$	plausibility $pl_{EO}(A)$
P{Detect}	0.40	0.568	0.40	0.60
P{Detect | Low}	0.30	0.50	0.30	0.50
P{Detect | High}	0.50	0.60	0.50	0.70

TABLE 4.1: Example for calculating belief and plausibility functions for perception node.

In the case of EEN, three more metrics are calculated (difference between $pl_E(A)$, $pl_O(A)$,$pl_{EO}(A)$ and *bel*). In this example, epistemic uncertainty increases for low performance of a LIDAR (Tab 4.1). The difference $pl_E - bel$ for *low* performance of LIDAR and $pl_E - bel$ for *high* performance of LIDAR corresponds to 0.10 decrease in the epistemic value. This may correspond to the *lack of knowledge* of the experts about the detectability of LIDAR at low performance. The SOTIF improvement measure may include better evidence and understanding on LIDAR detectability at low performance. The ontological uncertainty is represented by the *unknown* variable only. In this example it may represent the expert belief about the SOTIF relevant development lifecycle of the perception system.

4.7 Summary

SOTIF requires modelling of performance limitations under triggering conditions. Owing to the randomness and a general lack of knowledge about the occurrence distributions, existence and behaviour of elements, traditional safety analyses methods do not provide a complete representation for SOTIF analysis. In this chapter, a modelling approach based on CBN was put forward. The approach was further extended using EEN for modelling aleatory, epistemic and ontological uncertainty. The CBN structure and CBTs are both provided by the experts. EEN provides a modelling framework that models all types of uncertainties.

Epistemic uncertainty modelling provides a representation for generic lack of knowledge of the expert about the system model. This may represent data requirement, modelling process and domain knowledge of the expert. However, EN and EEN are advocated to represent subjective ignorance of the modeller in literature. This limits the learning techniques availability too. If data is used as the centre-piece of the analysis, DST based modelling techniques may be challenging to implement. A metric can be defined that intakes the discovered triggering conditions, understanding of confounding phenomena, collected data, known and unknown scenario discovered for a given ODD. This metric then can be used to model epistemic uncertainty values.

If data is used for structure [60] and CBTs determination [59], EN and EEN pose some limitation, as described in the section (Sec. 4.5). The epistemic uncertainty is then measured through a certain mathematical principle and represents a unique semantic of epistemic uncertainty in the model.

Ontological uncertainty representation in the presented framework is purely based on the expert guess. Generally, data may not correspond to the ontological uncertainty formalism i.e., they do not have *unknown* states in their labels.

5

Systematic Modelling, Estimation and Discovery of Perception Performance Limiting Triggering Conditions in Automated Driving

"God does not play dice with the universe."
– Albert Einstein, *German Theoretical Physicist*

In this chapter, a novel causal framework of SOTIF to model, estimate and discover triggering conditions relevant to selected performance limitations in automated driving is presented. The framework addresses the limitations of existing modelling tools discussed in the previous chapter (Ch. 3). In summary, the framework models the initial causal structure based on the expert knowledge and existing documented knowledge into a CBN and learns the CBTs from data. The resulting learnt CBN estimates the causal effect of triggering conditions using this model. Once the causal effect is estimated, the SOTIF modification can be formalized and implemented. Moreover, the model is also tested using test databases to provide an indication of novel triggering conditions in the scene. The novel triggering conditions are then refined into augmented causal models and causal effects are estimated again.

Sec. 5.1 introduces the chapter contributions of the causal framework while Sec. 5.2 provides a detailed overview of the framework. The last section (Sec. 5.3) provides the two representative algorithms of the framework.

5.1 Chapter Contribution

The overarching scope of the causal framework is to identify, model and quantify the emergence of performance limitations in the presence of triggering conditions influencing the scenarios. The concepts of causality and CBN to model the *aleatory* uncertainty in causal relation and occurrences of performance limitations and triggering conditions are utilized, while also utilizing iterative refinements and the concept of inferential statistics (e.g., p-value hypothesis testing), confidence interval and statistical variation to measure salient semantics of *epistemic* uncertainty. The framework provides a hybrid safety analysis approach; a unique provision of approach based on both expert knowledge and data driven engineering processes. In doing so, this dissertation provides solution on the following.

(S1) Measurement metrics and explanation of performance limitations

(S2) Relevant triggering conditions extraction

(S3) Convergence towards a manageable set of triggering conditions

(S4) Derivation of open context model from the SOTIF standpoint

(S5) Catalogue of abstract scenarios

(S6) Evaluation of the causal effect of one or more triggering conditions on performance limitations

(S7) Confidence building on the identified causal effect

(S8) Targeted SOTIF oriented modification of HAD function or ODD based on the causal effect calculations

(S9) An iterative SOTIF framework to assist semi-automated discovery of triggering condition

The contributions mentioned are represented purely from the SOTIF standpoint. The contributions also provide coverage and solutions to the limitations of the safety analysis methods discussed in the previous chapter. Understanding the relative frequency of performance limitation for a given scene is an indication of quantified risk for SOTIF [1]. These indicators can filter large datasets to build scenario catalogues to perform V&V. Evaluation of causal effect of triggering conditions on the performance limitation may explain the performance limitation occurrence in a scene. Every identified causal effect of triggering conditions on performance limitations increments the understanding of open context from the SOTIF standpoint, structuring the open context in this way. Moreover, since combinations of causal relations are also assumed, the framework also allows to identify, model and quantify complex causal structure in the open context.

Based on the identified causal structure, a finite scenario catalogue can be identified for providing a meaningful and manageable abstraction to the open context. The identified causal structure provides the elements to define necessary modification for SOTIF.

Identified causal effects may require indicators to provide reasonable confidence in the robustness of results. In addition, they may indicate further steps in case of lower level on confidence on the results.

One of the attributes of the open context is the presence of unknown triggering conditions that are also representative of unknown hazardous scenarios [2]. Unlike legacy systems where expert-based techniques have proven to work well, the open context nature requires an innovative solution to identify triggering conditions. The proposed framework provides an informed and systemic methodology to discover the triggering condition based on testable implications.

Some of the sections that constitute this chapter have been published as scientific contributions [1, 2].

5.2 Detailed Overview

In this section, a general overview of the proposed framework is provided as depicted in Fig. 5.1. The framework can be broadly divided into the following parts.

1. Knowledge acquisition

2. Databases
3. Parameter learning
4. Estimate and plausibilize
5. Explicate confidence
6. SOTIF improvement measures
7. Validate, refine and augment

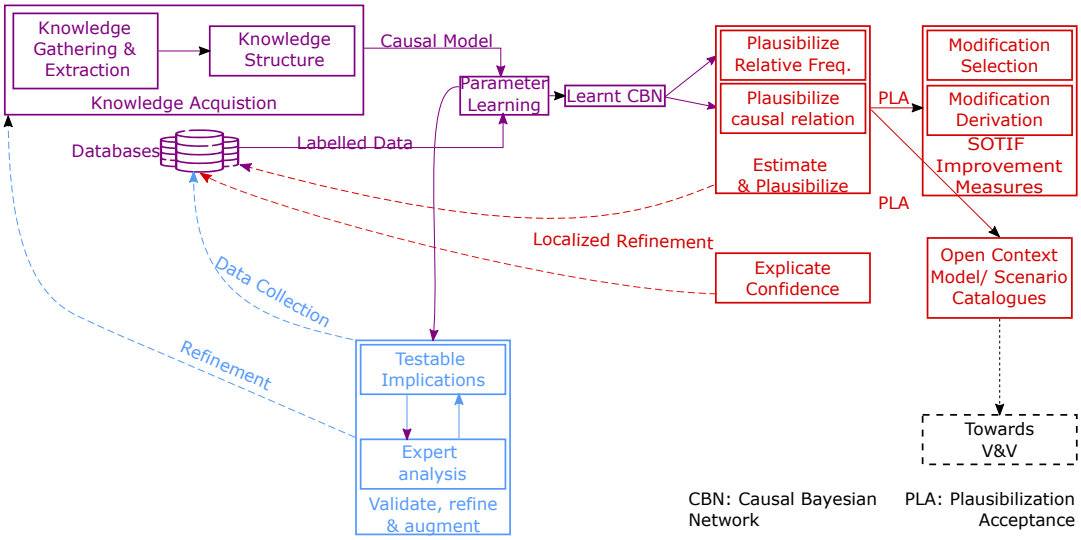


FIGURE 5.1: Detailed overview of the causal framework for SOTIF. The framework initiates with the knowledge acquisition that results in a causal model represented by a Causal Bayesian Network (CBN). With available data, parameter learning is performed on the CBN and causal effect of triggering conditions on the performance are estimated. Based on the causal effects, the SOTIF improvement measures can be formalized and implemented. The framework also supports multiple refinement iteration loops for data management. The CBN model is also tested against data and the results are used to identify novel triggering condition within the data. The novel triggering conditions are then modelled and estimated.

This section explains the overarching principles of each block. It also provides the methods and implementation schemes that can be used inside each block of the framework. Such implementation possibilities play a vital role in arguing the generalizability of the proposed framework. The implementation freedom of the framework with mathematical schemes provides an encompassing skeleton to analyse, evaluate, quantify and improve SOTIF. Overarching principles can be applied to diverse fields. However, this thesis only focuses on the information relevant to SOTIF.

5.2.1 Knowledge Acquisition

This block describes the collection of knowledge, information and data for the analysis and is depicted in Fig. 5.2. The focal point of this block is information gathering from various sources available about the system and its deployment context. Knowledge is extracted in the form of performance limitation and triggering condition from the gathered information. It is important to note that for each performance

limitation metric the existing knowledge on triggering conditions may vary. Moreover, it may also vary based on the chosen system and intended deployment context. Finally, a hypothesis on the structuring of the extracted knowledge is provided, resulting in the hypothesized causal model. This knowledge also stems from the same sources mentioned earlier. The prompted causal structure can be complex in nature. It may contain confounder, collider and forks sub structures [66, 88]. Moreover, triggering conditions may also have causal relations.

The knowledge acquisition process can be seen as the process of epistemology, thus reducing epistemic uncertainty about SOTIF. However, this step is analogous to other safety analyses techniques (e.g., FTA).

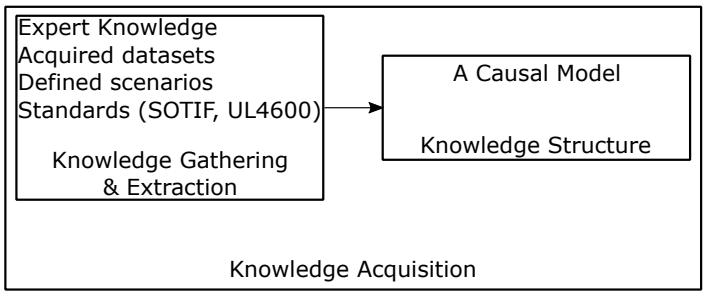


FIGURE 5.2: Detailed overview of the knowledge acquisition block. The block comprises of three process (1) Knowledge gathering & extraction: This process gathers knowledge about the system under development and deployed environment, the existing performance limitation of the system and potential triggering conditions in the environment. Knowledge extraction process extracts the relevant performance limitation and triggering conditions. (2) Knowledge Structure: This process models the extracted knowledge is a causal graph.

5.2.1.1 Knowledge Gathering

The knowledge gathering process includes but is not limited to expert knowledge, standards, acquired datasets and predefined scenarios.

Expert Knowledge

Expert knowledge has been the fundamental building block for any safety analysis technique [48]. Experts from the domain of the system and safety field provide their knowledge about the system. The process starts with definition of the analysis. In the case of HAD vehicles and SOTIF, this can be seen as the identification of performance limitations under triggering conditions. At this stage, reliability of experts can also be defined [39].

Acquired Datasets

Acquired datasets provide *facts* about the real world. In essence, they can serve as the first-hand representation of what can go wrong from the SOTIF standpoint.

Defined Scenario

To understand all the aspects of the analysis, the defined scenarios should be studied. This exercise further enhances the knowledge gathering process. Predefined scenarios increase explainability of the choices made in the knowledge gathering

process by enhancing the understanding of the driving context, the perception system and existing setup of the HAD vehicle. The abstraction at which these scenarios are defined may or may not govern the abstraction of knowledge gathering process. This depends on the data collection and human expert judgement.

Standards

The knowledge gathering process can be extended to normative standards. ISO 21448 [54], UL 4600 [123], BSI/ PAS 1883 [17] and ISO TR 4804 [56] are some of the standards related to HAD vehicle safety. It is worth mentioning here that the prime focus of this work is SOTIF and in that context all other standards are considered to augment the SOTIF standpoint.

5.2.1.2 Knowledge Extraction

Knowledge extraction process summarizes the extraction of required performance limitations and triggering conditions for a given analysis.

The prioritization of the performance limitation is important because performance limitation importance vary with the ODD and HAD vehicle function e.g., an FP probability may be deemed more important for Automated Emergency Braking (AEB), while for a HAD which also provides data to map making process, position trueness may be prioritized.

The prioritization of triggering conditions is important for the following reasons.

1. Triggering conditions selection is dependent on the performance limitation selection e.g., for an FN probability, occlusion is deemed more important than for an FP probability.

2. Triggering conditions selection is dependent on the availability of datasets.

3. Triggering conditions selection is also dependent on the HAD vehicle system under study e.g., a camera-based perception will result in a very different selection of triggering conditions than a LIDAR based perception.

5.2.1.3 Knowledge Structure - A Causal Model

The culmination of knowledge gathering and extraction process is the structure of the causal model. Knowledge structuring process induces the possible causal relations for the extracted performance limitations and triggering conditions. It utilizes similar resources as ones in the identification of triggering conditions for a performance limitation. In this framework, the resultant knowledge structure is causal in nature i.e., it assumes a causal connection between its various variables. The causal model is an interpretation of reality. Many representations of causal models exist including causal diagram, structural equation and logical equations [89]. CBN, a special case of causal diagram is used for representation. Moreover, the following assumption is also taken.

Assumption 6 *The causal model is the best representation of the open context and HAD vehicle performance (i.e., reality).*

Assumption 6 is seldom fulfilled as all models are at best an abstracted representation of reality [126]. Indicators to model this assumption assist in taking informed

decision about the causal model. CBN structure can be modelled based on the expert opinion as well as knowledge [37] or on structure learning through data [98]. However, the number of graph candidates grow exponentially based on the number of variables for structure learning techniques [60]. Moreover, if observational data is used for learning, discerning true graphs from other graphs that model the same set of conditional independence is also an intricate challenge. Owing to these reasons as well as challenges and the legacy of the safety analysis techniques, the initial CBN structure is based on expert knowledge solely.

5.2.2 Databases

Since one of the products of this framework is to provide plausibilized causal model for a performance limitation, causal relation can be explained by gathering evidence. Databases are gathered from the real world and simulations models. Databases are a representation of the chosen phenomena that occur in the real world. They can be either experimental or observational.

Data are the *facts* about reality [89]. Data constitutes the basic building block of knowledge and wisdom (Fig. 5.3). In general, data provides the basis of decision making and analysis of reality. They provide the smallest units of factual information about reality which then can be used to analyse, calculate and evaluate the reality for decision making. For example, rain = 5 millimetres (mm) is a measurement that constitutes datum. Data, information, knowledge and wisdom are closely related concepts with an increasing amount of certainty about reality.

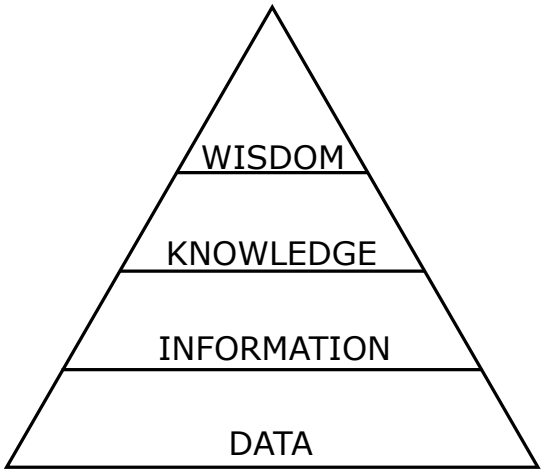


FIGURE 5.3: Hierarchy among data, information, knowledge and data. In this hierarchy, data represents raw values, information represents contextualization of data, knowledge provides structured information and wisdom represents the abstracted concept about reality.

Data without context is considered meaningless. In the context of region, 5mm of rain can give *information* that will vary between *low* or *heavy* rainfall based on the region. Structured and organized information based on cognitive processing becomes knowledge. This step is explained in the previous sections. Knowledge provides the answers to the *"how"* questions. Finally, wisdom is an extrapolative process which includes knowledge in ethical framework [26]. Databases, however, like all models of the world, are an abstract representation of reality. They are based on assumptions e.g., random samples, independence, normality etc. These assumptions should be considered in models based on data.

The collection of databases proposed in the framework is aligned with the guidelines provided in the SOTIF standard. The standard argues to identify known and unknown unsafe scenarios through V&V process [54]. It also advocates usage of test cases, scenarios and accident databases for the V&V activities. In this regard, identification and plausibilization of relevant triggering conditions through expert knowledge and databases can augment the probability of identification. Moreover, the databases should also be collected for cases where the causal effect of the triggering conditions is unknown [54].

The framework proposed in this dissertation thus provides unified artefacts to identification, V&V clauses of SOTIF.

5.2.3 Parameter Learning

Learning provides the basis of mathematical operations that can be used to evaluate and estimate various products of the framework i.e., estimation of causal relation and testing implication on test datasets for discrepancies between the *learnt* causal models and databases. Learning in this framework is the assimilation of causal structure and its representative data.

CBN is a graphical network that provides causal representation for probabilistic relations for a set of variables [88]. Parameter learning is the task to estimate all the conditional probabilities (i.e., defining probabilistic relations) from the datasets, given the causal structure [59]. From the dataset standpoint, the learning scheme can be divided into two categories.

1. Learning from complete data
2. Learning from incomplete data

5.2.3.1 Learning from Complete Data

Complete data indicates no missing values in the observations. Multiple learning techniques have been introduced in the literature for such type of datasets, two of those techniques are discussed here.

Maximum Likelihood Estimate

Maximum Likelihood Estimate (MLE) is a common strategy for parameter learning in BN. The underlying principle of the MLE is as follows: For any random observation $(O) \in O^1, O^2, \ldots, O^n$ over a set of random variables $X^1, X^2, \ldots, X^n$, the estimated value of $\hat{O}$ is based on the parameter θ, if it maximizes the value of the likelihood function $P(O|\theta)$. MLE is the most commonly implemented algorithm for parameter learning of BN.

Bayesian Methods

The underlying principle of the Bayesian method is as follows: For any unknown distribution and a dataset (O), θ is a random variable with a prior distribution $p(\theta)$; the observed probability, namely $p(\theta|O)$, can be estimated based to the prior knowledge or assumed distribution. The aim of this method is to calculate the posterior probability $p(\theta|O)$.

5.2.3.2 Learning from Incomplete Data

Incomplete data indicates missing values in the observations. Multiple learning techniques have been introduced in the literature for such types of datasets, three of those techniques are discussed here.

Expectation Maximization

The underlying principle of the expectation maximization is as follows: for incomplete observed data, the inference algorithm of BN can be used to predict the missing values of the dataset. This step renders the dataset complete. Expectation maximization includes two steps: (1) the initialization step: in this step θ is assigned a random value. (2) expectation calculation step.

Robust Bayesian Estimate

The underlying principle of the robust Bayesian estimate is as follows: Probability interval instead of traditional point probability estimate for each variable is provided. Robust Bayesian estimate differs the expectation maximization by not taking any assumption about the missing values while implementing parameter learning.

Monte Carlo Method

The underlying principle of the Monte-Carlo method is as follows: For any given function $f(X)$ of joint probability distribution $P(X)$ where X is the set of random variables, Gibbs sampling [43] is utilized to perform parameter learning.

The causal model may not be the exact representation of data and vice versa. This assumption is tested in the testable implication step (Sec. 5.2.7.1).

Assumption 7 *Databases distributions and causal models have one to one correspondence.*

This assumption asserts that a distribution P satisfies the independence defined in causal model G if P represents conditional probability distributions associated to G and vice versa. These concepts are termed as I-maps and D-maps [66].

Learning step (Fig, 5.1) is generally performed on the observational data as experimental data cannot be gathered for reasons discussed in Ch. 3.

5.2.4 Estimate and Plausibilize

Estimate and plausibilize block of the introduced framework provides two explanations.

1. Plausibilize relative frequency
2. Plausibilize causal relation

5.2.4.1 Plausibilize Relative Frequency

Relative frequency may provide occurrence of phenomena. Such measure can provide the quantification to SOTIF analysis under uncertainties. For example, the relative frequency of FN for LIDAR detection can be seen as a quantitative measure to LIDAR's performance limitation. Relative frequency can be donated by probability $P(.)$ of a random variable.

5.2.4.2 Plausibilize Causal Relation

Causal relation plausibilization provides two main explanations.

1. It explains the existence of causal relations.
2. It explains the underlying causal relation through a measured causal effect.

The so-called learnt causal model and a SOTIF relevant causal query forms the basis for estimation and plausibilization block. This block essentially answers the causal queries around SOTIF. Unlike FTA and STPA, this framework uniquely combines expert knowledge in the form causal model with the databases. The learnt causal model along with the causal query answers the causal effect of triggering conditions on the selected performance limitations e.g., the causal effect of occlusion on the FN in a specific setting. The results can be subjected to SOTIF improvement measures. They can be also subjected to localized refinement i.e., based on the results and expert knowledge; further data collection can be initiated.

Ideally, causal relations should be estimated through randomized controlled experiments [15]. However, these experiments may not be feasible as discussed in the previous chapter (Ch. 3). Alternatively, observational studies are used for prediction. Plausibilization of causal relations can be performed by virtue of mathematics defined by Judea Pearl [89]. Pearl's causal meta-model for such plausibilization involves a three level of abstraction, which he calls the ladder of causation [87]. These levels are as follows.

- Association calls for predictions based on the passive observations.
- Intervention calls for predictions based on the deliberate alterations of the environment and producing a desired outcome by choosing the right alterations.
- Counterfactual calls for predictions based on imagining the alterations of the environment and producing desired outcome by imaging the right alterations.

While implementation of associational and interventional queries is part of this dissertation, counterfactual queries are considered out of scope of this thesis.

Association

Association explains relevant relations by sensing the patterns in the input data of two variables. Pearl characterizes this phenomenon by the question "What if I see…?" For instance, imagine a safety analyst asking, "How likely is the FN/TP[1] probability of a LIDAR based detection to go up, given that the observer observes high occlusion phenomena?" Such queries identify and plausibilize triggering conditions and thus are SOTIF relevant. They can be answered by collecting and analysing data. In the example above, the question can be answered by first taking the data consisting of all the detections, selecting only with the high occlusion instances and then focusing on the FN/TP instances being true. This proportion is known as conditional probability and mathematically defined by the Bayes rule (Definition 11). It is important to note that the left-hand side of Eq. 4.2 is the building block of the equation defined for BN (Eq. 4.1). Conditional probability in general provides more meaningful causal relation and somewhat addresses the limitations of existing safety analyses modelling techniques discussed in the previous chapters.

[1] True Positive

The ability of conditional probability to model the randomized association between two random variables provides a representation to *aleatory* uncertainty.

Association may have evident causal interpretation or it may only show correlation [87]. Conditional probabilities may or may not have causal interpretation [87, 88]. In this thesis specifically, the causal treatment is associated with intervention.

Intervention

Intervention postulates distinct causal relations between events. In this level of causal queries, the world is altered. Considering the previous example, a typical question can be: "What will happen to FN probability, if the high occlusion occurrences are doubled?" This query cannot be answered by examining the history since it will alter the model of reality, which may have different reasons for FN probability to change.

The ideology of intervention goes hand in hand with the SOTIF analysis. The goal of SOTIF analysis is to analyse unsafe scenarios, thus a means to intervene on those scenarios can be provided through. Mathematically, these queries are answered by using Eq. 4.5.

Naturally, the identified causal relations represent the high-level model of the open context. They also provide input as scenario catalogues to the V&V clauses of the SOTIF. The established causal relation by virtue of causal effect calculation is tested for its robustness using mathematical indicators. It is also used as the basis for the required SOTIF improvement measures initiation.

5.2.4.3 Plausibilization Acceptance

Plausibilization acceptance indicates importance of the measured relative frequency (Sec. 5.2.4.1) or causal effect (Sec. 5.2.4.2). The acceptance can be based on an expert decision-making based on her experience or a reference value defined for each metric calculated in the previous step. Suppose, τ_1, τ_2 and define the allowable reference for probability and conditional probability (associational and interventional). Then $P(.) > \tau_1$ and $P(.|.) > \tau_2$ define the existence of potential hazardous behaviour pertaining to a performance limitation metric independently or under measured triggering condition(s).

5.2.4.4 Plausibilization Rejection

If the measured conditional probabilities and treatment effects are relatively smaller i.e., $P(.) < \tau_1$ and $P(.|.) < \tau_2$, the performance limitation can be considered negligible or triggering condition can be argued to have insignificant impact on the performance limitation of the system in the given context.

5.2.4.5 Localized Refinement

The localized refinement is triggered in cases where the human expert cannot decide between acceptance or rejection of the measured relative frequency or causal effect. In this case, more data is required to produce any decision. The localized refinement addresses the epistemic uncertainty about the causal relation by collecting more data[2].

[2]Data is the basic building block and can be converted to knowledge, thus decreasing *epistemic* uncertainty

5.2.4.6 Open Context Model/ Scenario Catalogue

The resulting causal structure post estimation and plausibilization can be referred to as high-level open context model. Such models not only increase the understanding of human experts about the nature of the context, but they also provide input for V&V of scenarios discussed in the previous chapter.

Just like the knowledge acquisition process (Sec. 5.2.1), this step can be seen as the process of epistemology, thus reducing epistemic uncertainty about SOTIF aspects of system model and its context.

5.2.5 Explicate Confidence

Evaluation of causal relations may not always provide the best estimates. Probabilities distribution of the databases are usually considered to be precisely known [61].

Assumption 8 *All the probabilities or probability distributions are known precisely.*

However, this assumption is seldom fulfilled. In order to service this assumption, several indicators can be used. These indicators measure indices of variation and statistical dispersion [38, 95, 90]. Confidence measurement block provides a representative method to provide the required metrics for robustness and confidence over the estimated results.

Confidence measures calculated through distribution represent the epistemic uncertainty about the datasets [88, 90]. If these measures are used on the association of two variables, they also represent the epistemic uncertainty about the SOTIF analyses and model which is based on the measured causal effect and the relevant dataset. Thus, it can be inferred that the measures represent the belief on the robustness of estimated and plausibilized causal relations. Some of the measures are given below.

5.2.5.1 Variation Ratio

The variation ratio is the simplest measure of statistical dispersion [38]. It is defined in terms of relative frequency mathematically as Eq. 5.1.

$$v = 1 - \frac{f}{N} \tag{5.1}$$

Where $\frac{f}{N}$ is the relative frequency of the occurrence of an event. For variation ratio to be zero, the relative frequency approaches unity. In terms of causal relation, such situation represents deterministic relations i.e., they can be represented with Boolean logic.

5.2.5.2 Information Entropy

Information entropy and its variants is the measure of uncertainty that is inherent in the variable's possible outcomes [106, 51]. For a random variable X with possible outcomes $x_1, x_2, \ldots, x_n$ and their probabilities $P(x_1), P(x_2), \ldots, P(x_n)$, the information entropy can be defined mathematically as Eq. 5.2.

$$H(X) = -\sum_{x=i}^{n} P(x_i) log P(x_i) \tag{5.2}$$

5.2.5.3 PPMI

Pointwise Positive Mutual Information (PPMI) is a measure of co-occurrence statistics. It measures the extent of occurrences of two events at random (or independent). In essence, the assumption is if the two events occur together more than expected, there exists a semantic relation between them. Mathematically, it is defined as Eq. 5.3 [95].

$$PPMI = 2^{log(\frac{p(y|x)}{p(y)})-(-log(p(x,y)))} \tag{5.3}$$

A PPMI value of 1 represents complete dependence of variables or events, while 0 represents independent events.

5.2.5.4 KL Divergence

KL divergence is a statistical distance that measures the difference between two probability distributions [68]. The mathematical notation is as follows (Eq. 5.4).

$$D(P|Q) = \sum_{i=1}^{n} P(x_i) log \frac{P(x_i)}{Q(x_i)} \tag{5.4}$$

Where P(.) and Q(.) are two different distributions of the same variable. Depending upon the distribution selection, many different results can be compared.

5.2.5.5 Standard Deviation

Standard deviation is the most common statistical dispersion measure. It defines the scatter of the expected value around the mean (Eq. 5.5).

$$\sigma = \sqrt{E[(X-\mu)^2]} \tag{5.5}$$

where σ is the standard deviation notation, $E(.|.)$ is the expected value and μ is the mean of the random variable. Pearl argues the equivalency of contingency sets and standard deviation as measures of the epistemic uncertainty modelled in the Dempster and Shafer theory [88]. In this manner, the dispersion measures can be viewed as an indicator of epistemic uncertainty.

5.2.6 SOTIF Improvement Measures

The implementation of ISO 21448 demands an iterative process to improve SOTIF. SOTIF improvement measure block of the framework supports clause 8 of the standard [54]. The estimated and plausibilized causal effects along with the confidence measures are used as input for this block. The activities address the SOTIF related risks. SOTIF advocates avoidance and mitigation measures to achieve SOTIF improvement measures. While avoidance represents elimination of risk, mitigation measures consider reducing risk where avoidance is not possible.

5.2.7 Validate, Refine and Augment

For the proposed framework to be representative, it must identify a manageable set of relevant phenomena to converge. This means that a systemic identification of all the relevant triggering conditions is required for a robust SOTIF analysis.

Scene modelling through the causal structure is based on a human expert knowledge, solely. This results in a fixed structure and comes at the cost of the best possible explanation of the scene given the dataset (Assumption 6). However, identification of unknown hazard scenarios (Fig. 3.8) and the triggering conditions thereof is a daunting task. Owing to the open context nature of the environment and lack of knowledge in general about the context and the system, modelling all the relevant triggering conditions based on the expert knowledge requires an iterative procedure as discussed in the SOTIF improvement measures.

The CBN structure with learnt CBTs together imposes implications e.g., "X and Y are dependent" or "knowing X can help us predict Y". However, if for newer datasets, these implications are tested and they fail, it can be assumed that the newer dataset may provide further input or information toward the predictability of Y in terms of a variable Z. This proposition forms the basis of validate, refine and augment block of the framework.

Novel triggering conditions can be identified through identification of scene anomalies in the datasets. In this way, a representative and manageable set of relevant triggering conditions can be discovered, identified, modelled and plausibilized, thus further improving the structure of the open context. Initial field tests can increase the databases, which in turn can assist in the identification of novel triggering conditions. Since the impact of the HAD vehicle on human traffic is significant, the growing databases need to be periodically checked against the current knowledge of the causal model of the open context.

Novel triggering conditions identification becomes more important for the HAD vehicle's safety deployed in the open context. Open context may also evolve over time and new phenomena may emerge, even if the initial triggering conditions set were sufficient. For example, the initial guess of an expert about the relevant triggering conditions for the LIDAR performance in terms of FN probability can be truncation, reflection and occlusion [1]. However, it is entirely possible that the FN probability is also influenced by the novel triggering conditions such as traffic density for the scenes the CBN was structured and of which the data was collected [2].

Validate, refine and augment block can be regarded as the process of epistemology, thus treating epistemic uncertainty. Although this step identifies rare events and thus rare triggering conditions, one can argue that the step also addresses the ontological uncertainty. However, in this dissertation the ontological uncertainty is purely defined as a subjective quantity proposed by the domain experts.

Validate, refine and augment block of the proposed framework introduces a methodology to identify, model and validate novel triggering conditions in a scene given a dataset to provide SOTIF assessment by subjecting the taken assumptions under test through testable implications and expert analysis.

5.2.7.1 Testable Implications

The learnt causal model i.e., CBN is based on the two main attributes.

1. A causal structure as a result of expert knowledge on the topic.
2. Probability distributions (e.g., conditional belief tables) resulting from datasets.

A causal relation present in the dataset may or may not be represented in the initial causal structure. The refinement and augment block takes the following assumption.

Assumption 9 *The novel relevant causal relations have traces of occurrence in the databases.*

Based on Assumption 9, if test datasets are used to evaluate and compare the initial CBN, it is possible to identify subsets of test datasets that involve traces of the occurrence of the novel causal relations with trigging conditions. The CBN can be augmented with novel triggering condition. In this regard, anomaly detection [6] and inferential statistics [108] based algorithm can be deployed. For example, p-value hypothesis testing is an inferential statistics technique. It is the probability of obtaining test results at least rare or rarer than the observed results. Mathematically, it can be given as.

$$p = P(T < t|H_0) \tag{5.6}$$

Eq. 5.6 shows the left tail p-value hypothesis testing. Similar notations can be derived for the right tail and two tail tests.

5.2.7.2 Expert Analysis

The identified subsets of data are then further analysed by an expert. Expert analysis results in the hypothesis of novel triggering conditions. The CBN is augmented using the novel triggering conditions, afterwards the resulting CBN can be estimated and plausibilized again. The need for expert analysis is aligned with the state-of-the-art safety analysis methods.

5.3 Representative Algorithms for the Framework

In the following sections, two representative algorithms supporting the framework are presented. The implementation division of the algorithms can be seen in Fig. 5.4. The framework is implemented in two different algorithms (with overlapping steps). This is because these algorithms address two different high level contributions i.e., estimation and plausibilization of causal relations as well as semi-automated discovery of novel triggering conditions. Since the framework provided is generic in certain aspects of its implementation and can be implemented in various mathematical languages, these algorithms provide a representative and implementation level facet of the framework. Especially the discovery of performance limiting triggering conditions through validate, refine and augment block require algorithmic level of understanding.

The common steps between the algorithms are explained once in the plausibilization algorithm (Fig. 5.5). Moreover, steps provided with ample detail and specificity within the framework explanation are not explained in the algorithms. Instead, the algorithms contain the explanation of more implementation-oriented artefacts.

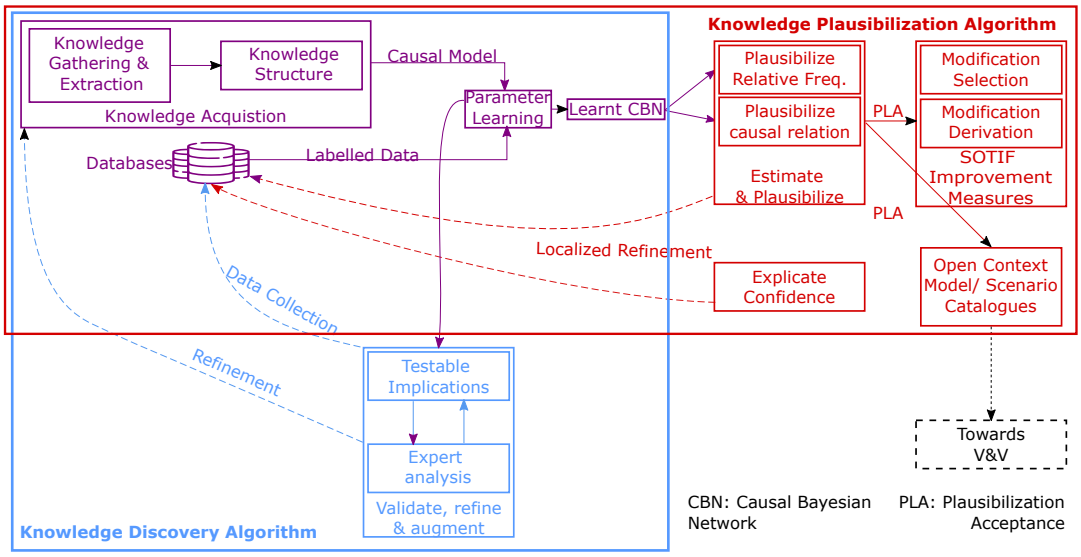


FIGURE 5.4: Algorithms proposed to implement the causal framework. The division of the framework into two different algorithms essentially addresses two different streams of high-level contributions i.e., estimation and plausibilization of causal effect as well as semi-automated discovery of novel triggering conditions. Evidently, both algorithms utilise similar initial learnt Causal Bayesian Network (CBN).

5.3.1 Knowledge Plausibilization Algorithm

A representative algorithm for the estimate and plausibilization iteration of the framework is provided in this section. The algorithm utilises CBN to identify, model and quantify performance limitations and triggering conditions present in a scene. The experts provide the initial CBN model. The CBTs are learnt from real world data.

Fig. 5.5 shows the flowchart of the algorithm. Estimate and plausibilize, explicate confidence and SOTIF measure steps of the algorithm (Fig. 5.5) are not explained. They have been already explained in the detailed overview of the framework (Sec. 5.2.4, Sec. 5.2.5 and Sec. 5.2.6). A detailed explanation of the remaining steps proposed in the flowchart (Fig. 5.5) follows.

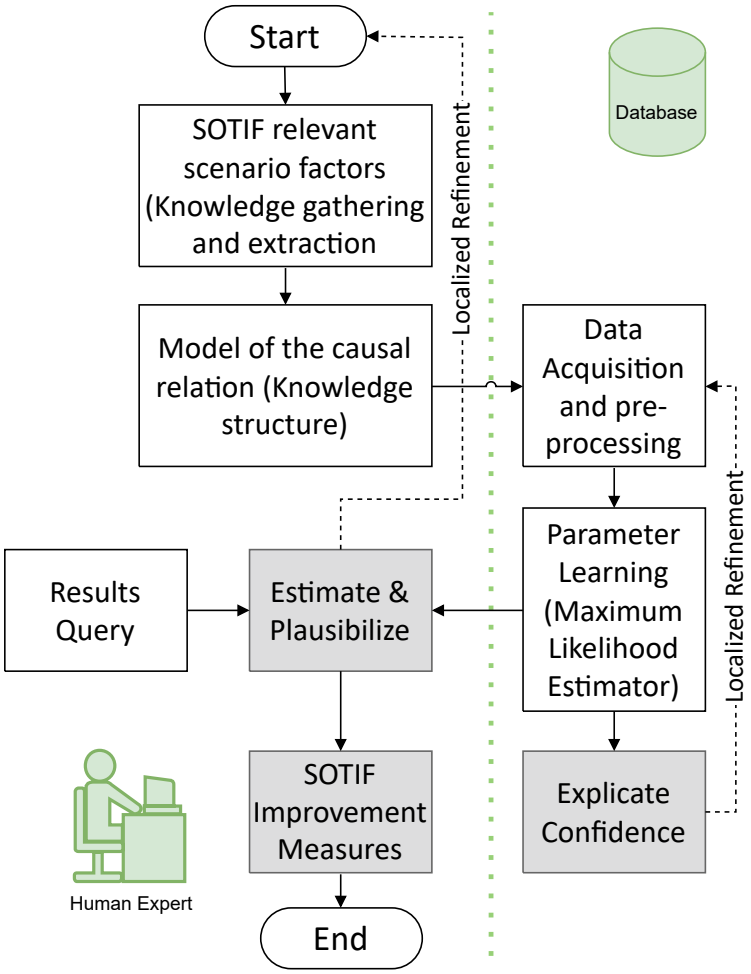


FIGURE 5.5: Flowchart describing the flow of the knowledge plausibilization methodology. SOTIF relevant scenario factors and expert knowledge are encoded into scene model defined by the Causal Bayesian Network (CBN) structure. Data is gathered accordingly and learning of parameters is performed.

5.3.1.1 SOTIF Relevant Scenario Factors

This step corresponds to knowledge gathering and extraction in the framework (Fig. 5.1). The first step towards modelling relevant SOTIF scenario factors is the identification of performance limitations and triggering conditions in a given scene [54]. SOTIF relevant scenario factors indicate triggering condition for performance limitations. Thus, to model SOTIF relevant scenario factors, identification of performance limitations and triggering conditions in a hypothetical scene is required. The identification process provides the implementable understanding of the knowledge acquisition process provided by the framework (Sec. 5.2.1). ISO 21448 provides a non-exhaustive scenery centric and dynamic element list of scenario factors along with some abstraction conceptualization of the scenarios [54]. Although this list can be a starting point, yet the knowledge acquisition process indicates that identification of triggering conditions for performance limitations should also include expert knowledge, pre-defined scenarios and acquired datasets for the purpose.

Different scenarios descriptions and system setups will yield different modelling factors based on the context of driving, perception system in question and existing setup, among others. For example, consider the following two descriptions.

1. Context: Highway, Perception: Radar based, Studied behaviour: False Positives.

2. Context: Urban, Perception: LIDAR based, Studied behaviour: Position Trueness.

Both descriptions may lead to different triggering conditions and performance limitations. In the former setup, the expert will be interested in tin cans, steel bridge and other such instance as these situations have results in RADAR interferences [125], while in the latter description, the expert may include weather conditions, exhaust gases and reflections.

In order to extract relevant scenario factors to identify triggering condition for performance limitations, the scenario factors from ISO 21448 [54] as well as expert opinion, previous data and constraints on data acquisition and/or data labels processes are considered (Fig. 5.6).

FP and FN detection may emanate SOTIF relevant undesired behaviour e.g., unintended braking of the HAD vehicle [54]. FN and FP can be modelled as performance limitations to assess SOTIF.

As an example, the LIDAR based perception system is used with the context of urban driving and position trueness, the following factors may be provided by the experts.

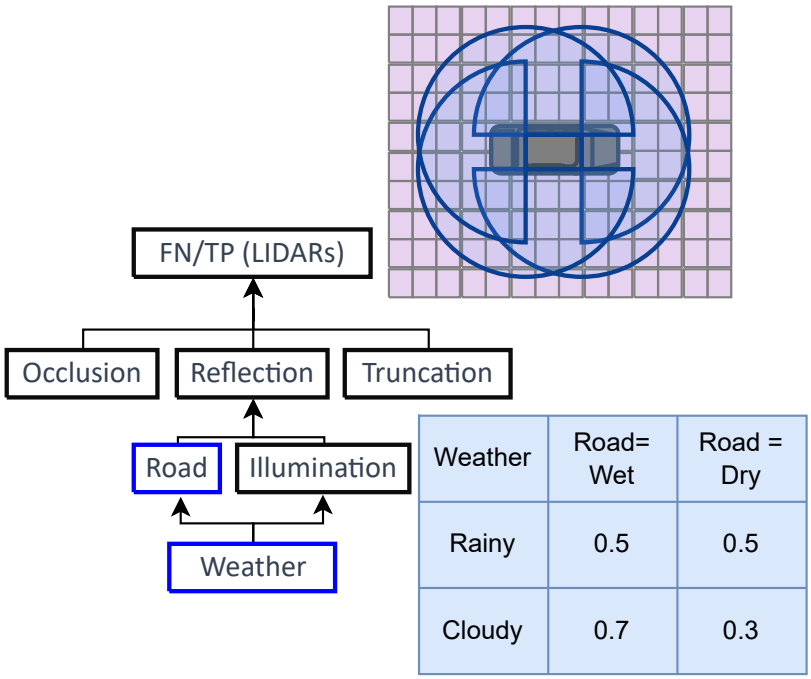
- **FN/FP**: The marginal distribution of the FN/FP.
- **Occlusion**: Relatively higher occlusion scenes can be observed with high probability since parked cars, road and environment infrastructure is believed to increase the probability of high occlusion scene.
- **Weather** conditions: LIDAR performance is affected by different weather conditions.
- **Reflection** from objects: The FN/ FP probability is affected by the reflections from objects (e.g., bus windows act as a mirror).
- **Illumination**: Reflection from the objects may increase from higher illumination.

The list of factors derived from the description above is non-exhaustive. Thus, the list should be updated iteratively as new knowledge through data, experiment and testing is acquired.

5.3.1.2 Model of the Causal Relation

This corresponds to the knowledge structuring step in the framework (Fig. 5.1). This step corresponds to a model that results from establishing causal relations amongst the triggering conditions and performance limitations. The resulting causal model may have complex causal relations. This step corresponds to the knowledge structuring step of the framework (Sec. 5.2.1.3).

Scenario description as provided in the previous section accounts for the nodes of the CBN. The domain experts initiate the derivation of the CBN structure by establishing the hierarchical dependencies between performance limitations and triggering conditions of the scene and provide propositions e.g., the proposition $p1$: high FN may result from highly occluded detections constituting a scene. Derived from



Weather	Road= Wet	Road = Dry
Rainy	0.5	0.5
Cloudy	0.7	0.3

FIGURE 5.6: An example of grid map and scene modelling attributed to the cells: LIDAR detections are discretized in grid cell around the field of view. Four LIDARs are attached to the roof of the HAD vehicle for detection. Bottom part shows a Bayesian network along with conditional belief table for $P(Road \mid Weather)$.

these propositions a CBN structure is constructed with edges between nodes representing the dependencies among performance limitation and triggering conditions. The proposition $p1$ may be represented as an explicit node (Fig. 5.6).

The CBN structure postulates that a parent node governs a child node by a causal mechanism which in turn is determined by the conditional probability distribution. The random attribute of the CBN model and the underlying causal mechanism also assists in modelling the aleatory uncertainty [1, 41].

5.3.1.3 Data Acquisition and Pre-processing

This algorithm utilises the datasets $\mathcal{D}$ that consist of fully observed instances of the nodes (complete data). This indicates that learning techniques described in Sec. 5.2.3.1 can be used.

$$\mathcal{D} = \xi[1] \ldots \xi[M] \tag{5.7}$$

Where M is the number of record instance of the dataset $\mathcal{D}$ and $\xi[.]$ represents a single instance of the data.

Based on the requirements imposed on the dataset by the CBN structure, new labels can also be calculated e.g., FN labels.

5.3.1.4 Parameter Learning

The CBTs are learnt once the CBN structure is established (Sec. 5.3.1.2) and the required dataset is available (Sec. 5.3.1.3). The CBTs determine the strength of the dependencies. In this algorithm, the MLE is used as the learning technique [66]. Moreover, non-parametric learning technique is used, assuming no prior probabilities.

For a variable X with its parents' variables $\mathbf{U}$, a parameter $\theta_{x|\mathbf{u}}$ for each combination of $x \in Val(X)$ and $\mathbf{u} \in Val(\mathbf{U})$ can be calculated. The likelihood function for such a case is as follows.

$$L_X(\theta_{X|U} : \mathcal{D}_{train}) = \prod_m \theta_{x[m]\mathbf{u}[m]} = \prod_{\mathbf{u} \in Val(\mathbf{U})} \prod_{x \in Val(X)} \theta_{x|\mathbf{u}}^{M_{train}[\mathbf{u},x]} \tag{5.8}$$

Where $\theta_{x|\mathbf{u}}$ is the learnt parameter and m represents the m^{th} instance in the dataset. The *train* and *test* subscript are used whenever *train* and *test datasets* are referred, respectively. The learnt parameter results from maximizing the likelihood function from Eq. 5.8.

$$\theta_{x|\mathbf{u}} = \frac{M_{train}[\mathbf{u}, x]}{M_{train}[\mathbf{u}]} \tag{5.9}$$

Here $M_{train}[\mathbf{u}, x]$ represents the combined occurrence of u and x. Eq. 5.9 defines the MLE.

5.3.1.5 Result Query

To perform SOTIF analysis, result queries are provided by the expert. For example, estimation and plausibilization of occlusion's effect on FN is a desired query. Ch. 7 provides details of multiple such queries.

5.3.1.6 Localized Refinement

The localised refinement step aims to provide improvement in the CBN (both structure and CBTs), to provide a sufficiently complete and exhaustive CBN model for SOTIF analysis. The hybrid approach presented here involves both expert knowledge and dataset, while also partially automating the approach may produce better safety models. Every step discussed in the algorithm and represented by the flowchart (Fig. 5.5) is subject to refinement iteratively, based on the analysis of the results. This includes addition/ deletion of triggering conditions, restructuring the CBN structure and/or acquisition of more data.

The term "localised" is deliberately used to differentiate this step from the semi-automated discovery of triggering conditions-based refinement of the CBN model, which is discussed in the next algorithm.

5.3.2 Knowledge Discovery Algorithm

An algorithm utilising the testable implications is proposed here. The algorithm tests the hypothesis of causal relations in a scene with respect to the initially proposed CBN in order to discover triggering conditions. The identified potential triggering conditions are then modelled, quantified and verified. The model of the causal relations and parameter learning steps are not discussed here as they overlap the previous algorithm implementation and are fully discussed in Sec. 5.3.1. Hypothesis tests on the learnt CBN using p-values statistics are performed [108, 75, 76]. The testable implications are the causal relations of the initial CBN as calculated by the probability distributions and defined per scene. Thus, the implementation results in subset of relevant scenes. These scenes are then analysed by the experts, and they provide refinement strategies, accordingly.

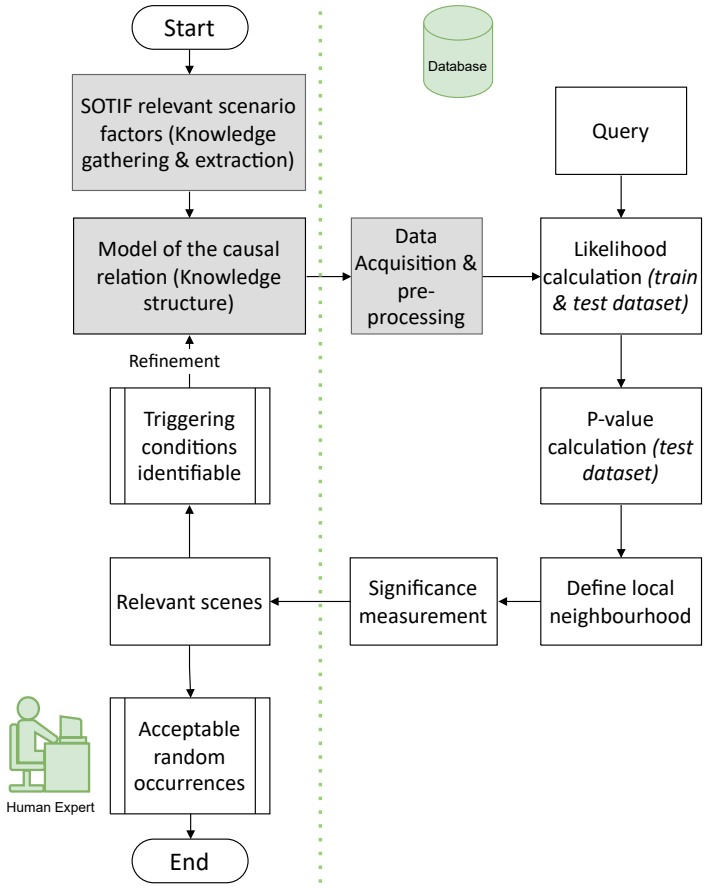


FIGURE 5.7: Flowchart describing the flow of the knowledge discovery methodology. SOTIF relevant scenario factors and expert knowledge are encoded into scene model defined by the Causal Bayesian Network (CBN) structure. The shaded steps are part of the previous publication [1]. Established Conditional Belief Tables (CBTs) (after parameter learning) are tested with p-value hypothesis and relevant scenarios are extracted, refinement steps are introduced and retested till a sufficiently accurate CBN is achieved.

5.3.2.1 Query

The query defines the causal implication defined in the CBN that needs to be tested. For example, if an FN relative frequency is predicted by two variables *Occlusion* and *truncation* (FN has two parent nodes in a CBN with Markovian conditions assumed), these variables together form a query of the causal implication.

5.3.2.2 Conditional Belief Likelihood Assignment

Given the conditional dependencies defined in the CBN structure (Sec. 5.3.1.2) along with the learnt CBTs (Sec. 5.3.1.4), for each test query the Conditional Belief Likelihood (CBL) is calculated as follows.

$$CBL^{j}_{x|\mathbf{u}} = \xi[j]_{x|\mathbf{u}} = \theta_{x|\mathbf{u}} \tag{5.10}$$

In the Eq. 5.10, $j \in M$ (train and test dataset), $CBL^{j}_{x|\mathbf{u}}$ or $\xi[j]_{x|\mathbf{u}}$ refers to realization of the random variable (node) x given realization of random variables modelled as its parents $\mathbf{u}$ in the jth data instance. For example, if $x : FN(Yes)$ given its parents nodes $\mathbf{u} : Truncation\ (Yes), Reflection\ (Yes), Occlusion\ (Largely\ Occluded)$ corresponds to jth row, then $CBL^{j}_{x|\mathbf{u}}$ assignment corresponds to $\theta_{x|\mathbf{u}}$.

5.3.2.3 P-values Calculation

Null hypothesis testing e.g., p-value calculation, has been used in testing CBN patterns [108, 75, 76]. The p-value hypothesis testing can be defined as the probability of acquiring test results that are rarer or at least equally rare than the observed (training data) results. In order to handle ties in the conditional probabilities (CBLs), ranges in the p-values have been proposed in the literature [75, 76]. The p-value testing ranges calculate the relative frequency of equally rare or rarer CBL in the train dataset.

$$M^{CBL^{j}_{x|\mathbf{u}}}_{lower} = \sum_{k \in \mathcal{D}_{train}} I(CBL^{k}_{x|\mathbf{u}} < CBL^{j}_{x|\mathbf{u}}) \tag{5.11}$$

$$M^{CBL^{j}_{x|\mathbf{u}}}_{equal} = \sum_{k \in \mathcal{D}_{train}} I(CBL^{k}_{x|\mathbf{u}} = CBL^{j}_{x|\mathbf{u}}) \tag{5.12}$$

Consequently, the p-value ranges can be defined as.

$$\begin{aligned} p^{j}_{x|\mathbf{u}} &= [p_{min}(p^{j}_{x|\mathbf{u}}), p_{max}(p^{j}_{x|\mathbf{u}})] \\ &= \left[\frac{M^{CBL^{j}_{x|\mathbf{u}}}_{lower}}{M_{train}+1}, \frac{M^{CBL^{j}_{x|\mathbf{u}}}_{lower} + M^{CBL^{j}_{x|\mathbf{u}}}_{equal} + 1}{M_{train}+1} \right] \end{aligned} \tag{5.13}$$

Where M_{train} corresponds to number of training data instances.

5.3.2.4 Significance Calculation

Significance is calculated at level α for p-values, to test the hypothesis. Given a distinct p-value a measure that quantifies the significance at level α can be written mathematically as follows [75].

$$n_{\alpha}(p) = I(p \leq \alpha) \tag{5.14}$$

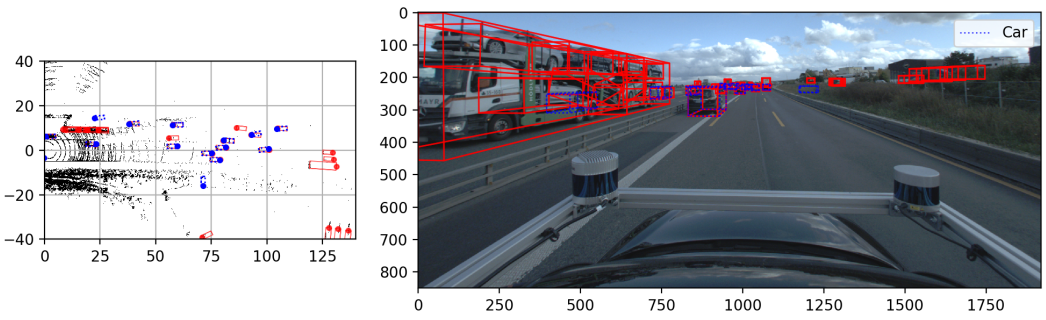


FIGURE 5.8: Scene representing the causal relations *"cars loaded on a trailer"*, *"ground truth labelling errors"* and *"vehicle activity"*.

Eq. 5.14 can be extended to p-value ranges and can be written as follows.

$$n_\alpha(p^j_{x|\mathbf{u}}) = \begin{cases} 0 & \text{if } p_{min}(p^j_{x|\mathbf{u}}) > \alpha \\ 1 & \text{if } p_{max}(p^j_{x|\mathbf{u}}) < \alpha \\ \frac{\alpha - p_{min}(p^j_{x|\mathbf{u}})}{p_{max}(p^j_{x|\mathbf{u}}) - p_{min}(p^j_{x|\mathbf{u}})} & \text{otherwise} \end{cases} \tag{5.15}$$

5.3.2.5 Local Neighbourhood Definition

Local neighbourhood is defined in the dataset by the distance-based methods traditionally [75] e.g., Euclidean distance can be used for continuous variable data instances, while Jaccard index can be used for categorical variable data instances.

In this algorithm, however, a novel scene level local neighbourhood definition is introduced i.e., the scene data instances in the test dataset M_{test} are combined into a single local neighbourhood. A camera-based image equivalent to such a scene is shown in Fig. 5.8. The red bounding box annotation depicts ground truth data instance, while the blue bounding box annotation depicts detection data instances.

The selection of scene as the local neighbourhood emphasizes that at scene level of abstraction, the data instances in the test datasets can be as different as the significance level α from the train dataset.

5.3.2.6 Relevant Scene Identification

Once the local neighbourhood is conceptualized, the relevant scene can be analysed by using the mathematical equations for a scene S defined as a local neighbourhood as follows.

$$N_\alpha(S) = \sum n_\alpha(p^j_{x|\mathbf{u}}) \tag{5.16}$$

$$N(S) = \sum I \tag{5.17}$$

Local neighbourhood scenes S that are considered as relevant scenes must satisfy the inequality $N_\alpha(S) > \alpha N(S)$. These scenes need to be further analysed.

The calculation steps defined so far can be produced for any combination of nodes in the CBN structure. However, in this algorithm, the calculation is limited to a single node of the initial CBN for an iteration of the algorithm.

5.3.2.7 Relevant Scene Causal Relation

Every relevant scene identified through hypothesis testing is subject to expert analysis. In the proposed algorithm, the relevant scene is assessed by the experts under two probable explanations.

Acceptable Random Occurrences

The first explanation relates the identified scenes as random occurrences i.e., experts may propose that identified scenes are random occurrences and no novel triggering condition is identifiable.

Novel Triggering Condition Identifiable

The second explanation relates the identified scenes as occurrences in which novel triggering conditions are identifiable i.e., the experts may propose relevant triggering conditions that should be taken into account in the CBN to assess SOTIF. For example, during analysis the expert may find scenes in which $traffic\ density$ may influence the FN probability, thus should be modelled, estimated and plausibilized as triggering condition in the CBN.

5.3.2.8 Refinement

Refinement step prescribes modelling, estimating and plausibilizing the identified potential triggering conditions (Sec. 5.3.2.7) into the CBN structure. A variable can be modelled into a CBN by virtue of four possible edge trails Koller et al. [66] to model a variable into a CBN. Only direct causal edge trail and confounding causal edge trail is considered in this algorithm to model the novel triggering conditions (Fig. 5.9).

Apart from new edge trail modelling, the initial CBN is also tested by removing an existing triggering condition.

Direct Causal Edge Trail

Direct Causal Edge Trail (DCET) is the simplest mechanism to model the novel triggering condition. In this modelling mechanism, the novel triggering condition directly effects a node in the existing CBN (Fig. 5.9(a)).

Confounding Causal Edge Trail

Confounding Causal Edge Trail (CCET) is based on the mechanism where a variable influences both the dependent and independent variable, causing a spurious association (Fig. 5.9(b)). In this modelling mechanism, the novel triggering condition controls two directly connected variables in the existing BN, simultaneously.

DCET and CCET calculation produce a similar number of relevant scenes for a given variable in a hypothesis test, as the number of parents for the variable remains similar for DCET and CCET. However, if the p-value hypothesis test for a novel triggering condition is performed for both dependent (child node) and independent (parent node) variable, and a DCET can be established for both variables, the refinement step may indicate a CCET i.e., occurrence of a confounding phenomenon. For example, consider FN has a parent node occlusion and after the p-value test as well as expert analysis, traffic density is proposed as the novel triggering condition. In

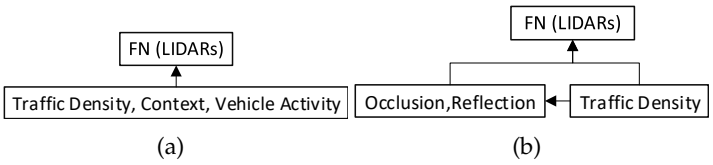


FIGURE 5.9: Refinement steps considered in this dissertation in the knowledge discovery algorithm. (a) Direct Causal Edge Trail (DCET) (b) Confounding Causal Edge Trail (CCET)

order to establish a CCET, individual DCET refinement should be validated for both FN and occlusion.

Triggering Condition Removal

Triggering condition removal steps challenges the initial CBN proposed by the experts (Sec. 5.3.1.2). Some of the variables modelled in the initial CBN may not conform to the p-value hypothesis test. For example, a variable truncation may initially be considered as a relevant triggering condition by the experts may not be a relevant triggering condition given the test datasets.

5.3.2.9 Validation

Once the refinement mechanism is decided and implemented, the CBN can be evaluated against the Relevant Scene Score[3] (RSS) before and after the adjustment concluded in the refinement step. The validation step includes a $proposition_{NTC}$ for the novel triggering condition from an algorithmic standpoint and final conclusion from the expert's standpoint as described in Algorithm 1.

Algorithm 1 Validation Algorithm Flow

if $(RSS^{node}_{initial} > RSS^{node}_{after})$ **then**
  $proposition_{NTC} = valid$
  $Expert\ Conclusion =$
  $Accepted\ Proposition$ **or** $Inconclusive\ Evidence$
else if $(RSS^{node}_{initial} < RSS^{node}_{after})$ **then**
  $proposition_{NTC} = invalid$
  $Expert\ Conclusion =$
  $Rejected\ Proposition$ **or** $Inconclusive\ Evidence$
end if

Where NTC is the novel triggering condition, $RSS^{node}_{initial}$ is the relevant scene score before the modification in the CBN and RSS^{node}_{after} is the relevant scene score after the modification in the CBN, relative to existing CBN *node*. The important aspect of Algorithm 1 is that the expert may or may not accept the valid proposition for a novel triggering condition $proposition_{NTC}$. The expert may evaluate a proposition and accept it based on the difference between $RSS^{node}_{initial}$ and RSS^{node}_{after}, past experience and representativeness of data etc. Novel triggering conditions that are finally deemed important and accepted by the experts are included in the CBN. The decision criteria described here significantly differs from the purely data driven techniques, where

[3]Relevant Scene Score is the number of identified scenes.

the decision is based on the results of the algorithm and no expert knowledge is taken into consideration.

6

Case Study and Implementation

"Somewhere, something incredible is waiting to be known."
– Carl Sagan, *American Astronomer*

In this chapter, the representation of real-world case study along with implementation details of the methodology developed in the previous chapter is provided. The chapter starts with the description of the case study (Sec. 6.1). Implementation details of the framework developed in the previous chapter are provided in the next section (Sec. 6.2).

6.1 Case Study

The case study consisted of a test vehicle on which a LIDAR based perception system was mounted (similar to Fig. 5.6 schematics). Two LIDAR experts and two safety researchers were involved in the execution of knowledge gathering process based on the expert opinion and literature.

6.1.1 Experimental Setup

The experimental setup for data collection consists of two Hesai Pandar 64 and two Velodyne Ultra Puck VLP-32C LIDAR sensors as part of perception system. The sensors are installed on the roof corners of a car. The collected and labelled data consists of bounding boxes, detection pose, visibility state and vehicle activity among others surrounding 360° of the HAD vehicle. A DNN is trained and used to detect cars.

Two separate labelled datasets correspond to detection and ground truth instances. These instances are labelled as a blue and red bounding box (Fig. 5.8). Most of the data was collected on different highways in Europe. However, part of the collected data also belongs to urban roads. Roughly twenty thousand labelled instances are available in both datasets. Two experts provide their opinions on LIDAR insufficiencies, triggering conditions and limitations.

6.1.2 Data Representation

In order to fully understand the effects of the triggering conditions, scenario factors and performance limitation around the vehicle, the spatial distribution of detections are discretized into a grid map (Fig. 5.6). Grid maps like discretization of the spatial distribution is important for the following reasons.

1. Relevant triggering conditions and scenario factors are spatially distributed e.g., for certain perception system dense fog will spatially effect the FN probability distribution.

2. Safety criticality around the HAD vehicle is variable in nature i.e., events' occurrence nearer to the HAD vehicle can be considered more critical generally.

Discretization around the HAD vehicle also leads to spatial association of the data instances at the respective detection points in space. Discretization of the grid map is based on the type of the coordinate system (e.g., polar or Cartesian) and grid size. A distinct CBN and CBTs based on the allocated data then represent a grid cell (Fig. 5.6). The CBN structure is kept constant in this implementation.

Suppose the data instances are distributed into $\mathcal{N}$ number of grid cells (thus $\mathcal{N}$ number of CBNs) based on the Cartesian (x, y) or polar (r, θ) coordinates of detection. The equations (Eq. 5.7, 5.8, 5.9) change to the following.

$$\mathcal{D}^k = \xi^k[1] \dots \xi^k[M^k] \forall k \in \mathcal{K} \tag{6.1}$$

Where $\mathcal{K}$ is a set as follows.

$$\mathcal{K} = \{1, 2, \dots, \mathcal{N}\} \tag{6.2}$$

Here k represents k^{th} grid cell and CBN. The likelihood function and learnt parameter can be rewritten as.

$$L_X(\theta^k_{X|U} : \mathcal{D}^k_{train}) = \prod_m \theta^k_{x[m]\mathbf{u}[m]} = \prod_{\mathbf{u} \in Val(\mathbf{U})} \prod_{x \in Val(X)} {\theta^k_{x|\mathbf{u}}}^{M^k_{train}[\mathbf{u},x]} \tag{6.3}$$

$$\theta^k_{x|\mathbf{u}} = \frac{M^k_{train}[\mathbf{u}, x]}{M^k_{train}[\mathbf{u}]} \tag{6.4}$$

As the representation occurs at the estimate and plausibilize block (Sec. 5.2.4), the above equations are only used for plausibilization algorithm. For semi-automated discovery algorithm, no data discretization is performed and equation defined in the previous chapter (Eq. 5.9) is used.

For the implementation necessary for this dissertation and based on the data availability for grid cells and completeness in the representation of each node of the CBN (Fig. 6.1), $x = 20$ and $y = 10$ meters were selected as the grid cell dimensions.

6.1.3 Data Collection and Annotation

Dataset collected consists of multiple label annotations. Selected variables are deemed important from the SOTIF standpoint. Moreover, only selected variables are discussed in the following.

6.1.3.1 FN

FN is defined as miss-detection i.e., an object presents in the ground truth dataset but not detected. FN has been advocated as a performance limitation metric by the SOTIF standard [54]. FN is defined as a binary variable with *"yes"* indicating occurrence of FN in an instance and *"no"* indicating the otherwise. In this thesis, FN represents the miss-detection of a car on the road.

6.1.3.2 Truncation

Truncation describes an object partially outside the field of view of the sensor. Truncation condition is defined as a binary state variable. Truncation is represented by two states; *"yes"* and *"no"*. Truncation may define a scene level condition[1].

[1]Scene level condition is assumed to remain the same across the spatial description of scene.

6.1.3.3 Reflection

Reflection describes reflective effects from road. Reflection condition is defined as a binary state variable. Reflection is represented by two states; *"yes"* and *"no"*. Reflection may define a scene level condition.

6.1.3.4 Occlusion

Occlusion is the effect of one object in a 3-D space blocking another object from view. Occlusion condition is defined as a multi-state variable. Occlusion is represented by four states; *"fully visible"*, *"partly occluded"*, *"highly occluded"* and *"unknown"*. Occlusion may define a scene level condition.

6.1.3.5 Illumination

Illumination describes the lightening conditions of the scene. Illumination condition is defined as a multi-state variable. Illumination is represented by four states; *"low light"*, *"day"*, *"night"* and *"tunnel"*. Illumination may define a scene level condition.

6.1.3.6 Road

Road condition is defined as a binary state variable. The road is represented by two states; *"wet"* and *"dry"*. Road may define a scene level condition.

6.1.3.7 Weather

Weather defines the current weather in the scene. Weather conditions are defined as a multi-state variable. Weather is represented by five states; *"cloudy"*, *"sunny"*, *"rainy"*, *"clear"* and *"not defined"*. Weather may define a scene level condition.

6.1.3.8 Context

The context of driving describes different type of roads in different scenes. Context condition is defined as a binary state variable. Context is represented by 2 states; *"highway"* and *"urban"*. Context may define a scene level condition.

6.1.3.9 Vehicle Activity

Vehicle activity is defined as a multi-state variable. Vehicle activity is represented by four states; *"parked"*, *"stopped"*, *"moving"* and *"others"*. Vehicle activity may define an instance level condition[2].

6.1.3.10 Traffic Density

Traffic density defines the amount of traffic on the road in a scene. Traffic density is defined as a multi-state variable. Traffic density is represented by three states; *"low"*, *"medium"* and *"very high"*. Traffic density may define a higher abstraction than a single data instance i.e., as a scene level condition. It may not be intuitive for a data instance of detection similar to other nodes such as weather, however, it defines the class of scene the instance belongs to.

[2]Instance level condition is assumed to change across the individual object in a scene.

6.2 Implementation

The framework discussed is implemented through the representative algorithms discussed in the previous chapter. Steps not included in the implementation are explained in the evaluation and results of implementation.

6.2.1 Knowledge Plausibilization Algorithm Implementation

In this section, the implementation of the plausibilization algorithm is provided for the LIDAR sensing dataset.

6.2.1.1 SOTIF Relevant Scenario Factors

Scenario factors that may effect the LIDAR system performance are provided by the experts (Sec. 5.3.1.1). As described in the knowledge acquisition block (Sec. 5.2.1), SOTIF scenario factors [54], expert inputs and available data form the basis of the inclusion of variables (Fig. 6.1). The following conclusions can be made.

1. Occlusion and truncation of objects may only generate scattered point clouds [82].
2. Various weather conditions may affect the road profile and light intensity. This may in turn result in reflections.
3. FN is used to represent the performance limitation of the LIDAR perception system, advocated by the SOTIF standard as an adequate measure [54].

6.2.1.2 Model of the Causal Relation

The nodes in Sec. 6.2.1.1 form the basis of the CBN structure. The nodes can be encoded into a CBN structure by using the following simple propositions.

Proposition 1

FN may be influenced by **truncation**, **reflection** and **occlusion** in detections.

Proposition 2

Reflection may be influenced by **road** conditions and **illumination**.

Proposition 3

Road conditions and scene **illumination** are influenced by **weather conditions**.
The proposition when encoded in a CBN structure results in Fig. 6.1. The CBN structure contains seven nodes.

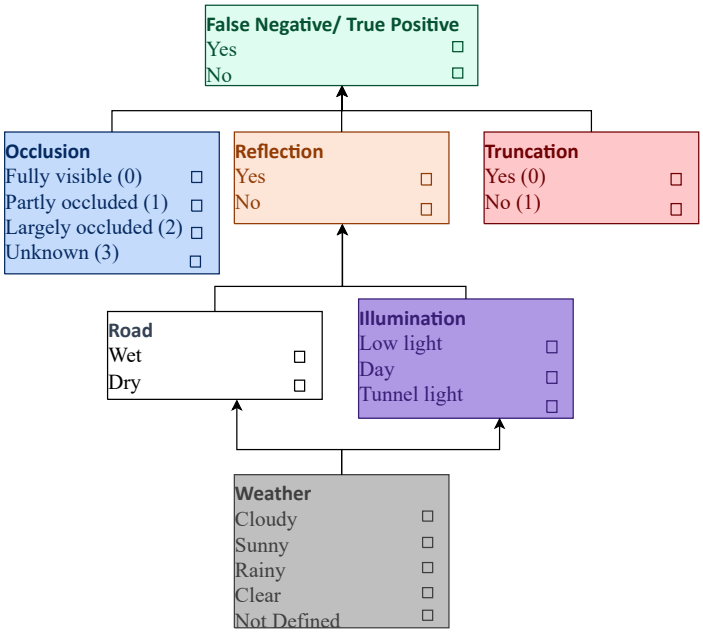


FIGURE 6.1: Causal Bayesian Network based on the SOTIF relevant scenario factors and expert knowledge describing the causal structure used in the implementation.

6.2.1.3 Data Acquisition and Pre-processing

All the relevant nodes of the CBN structure, except FN, are labelled (Fig. 6.1). In order to evaluate FN for each data instance, Mean Squared Error (MSE) is used.

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(Y_i - \hat{Y}_i)^2 \tag{6.5}$$

Where n represents number of samples, Y_i represents the ground truth and $\hat{Y}_i$ represents the detection. Eq. 6.5 is executed for individual detections while tracing a corresponding sample in the ground truth using x and y values, where (x, y) define the centre of the bounding box for a detection and ground truth data.

The philosophy behind the calculation lies on the assumption that data instances present in both detection and ground truth sets with a small margin of error η are considered TP, while data instances only present in ground truth set are considered FN and so on. Moreover, based on the defined optimal range for LIDARs, spatial cut off values are defined at $|x| > 140$ meters and $|y| > 50$ meters.

6.2.1.4 Parameter Learning

Parameter learning is performed using MLE (Eq. 6.3). Parameter learning for individual CBN (representing a grid cell) using its corresponding data instances and Eq. 6.3 is implemented.

6.2.1.5 Localized Refinement

The localized refinement steps are discussed in the results (Ch. 7).

6.2.2 Knowledge Discovery Algorithm Implementation

In this section, the application of the methodology on the LIDAR sensing dataset explained in the previous section is demonstrated.

Sec. 5.3.2.2 and Sec. 5.3.2.3 are not discussed as they are purely mathematical calculations. Moreover, Sec. 5.3.2.6, Sec. 5.3.2.7, Sec. 5.3.2.8 and Sec. 5.3.2.9 are discussed as part of the result section (Sec. 7).

6.2.2.1 Date Acquisition and Pre-processing

This step acquires the same dataset used in the plausibilization algorithm. However, the datasets are not discretized in this step as grid maps. Moreover, a randomized division of train and test datasets (80% and 20%) is performed multiple times to perform parameter learning using equation Eq. 5.9.

6.2.2.2 Parameter Learning

Parameter learning is performed using MLE (Eq. 5.9). The parameter learning differs from the one explained in Sec. 6.2.1.4 as different equations are used.

6.2.2.3 Significance Calculation

The significance level α is chosen at 5% to perform the significance calculation. The choice is made purely on the premise that in the state-of-the-art implementation of p-value hypothesis testing, α is chosen at this level [75].

6.2.2.4 Local Neighbourhood Definition

Local neighbourhood in the test datasets is defined based on the scene categorization. This is a novel approach to present local neighbourhood as distance-based subsets of dataset instances are clustered together traditionally. Such definition equips the implementation with the possibility of identification of the relevant scenes which then can be subject to refinement (Sec. 7.2.1). Fig. 5.8 depicts one such scene.

7

Case Study's Results

"What we know here is very little, but what we are ignorant of is immense."
– Pierre Laplace, *French Polymath*

In this chapter, results from the implementation of the algorithms are presented. Sec. 7.1 provides the results of estimation and plausibilization process. Sec. 7.2 summarizes the results of the discovery of the perception performance limiting triggering conditions using p-value hypothesis testing. Sec. 7.3 discusses the results of the second iteration of the estimation and plausibilization process. Finally, chapter summary is provided in Sec. 7.4.

7.1 Knowledge Plausibilization Algorithm Results: First Iteration

In this section, the results of the implementation are presented. The relevant *query* (Fig. 5.5) is explained under the following features.

- **Performance Limitation Maps:** Performance Limitation Maps (PLMs) represent the marginal posterior probability distribution $P(x)$. The naming is intentionally done in this manner in order to provide a SOTIF-oriented semantics to the result.
- **Conditional Performance Limitation Maps:** Conditional Performance Limitation Maps (CPLMs) represent the conditional probability (both associational and interventional) i.e., $P(y|x)$ or $P(y|do(x))$. Interventional calculation becomes important in places where confounding phenomena is present. In the light of ISO 21448, it can be seen as how triggering conditions influence the performance [54].
- **Explicate Confidence:** Explicate confidence metric represents the measurement confidence of different queries. In the results, PPMI is used. PPMI can be an indicator of epistemic uncertainty in the causal relation. A smaller value indicates a higher epistemic uncertainty.

Based on the above-mentioned quantities, SOTIF improvement measures, analysis conclusions and localized refinements are derived. In the following, some of the most important queries are presented and discussed in detail.

7.1.1 False Negative

7.1.1.1 PLM

PLM for FN is shown in Fig. 7.1. The following evident analysis conclusions can be drawn.

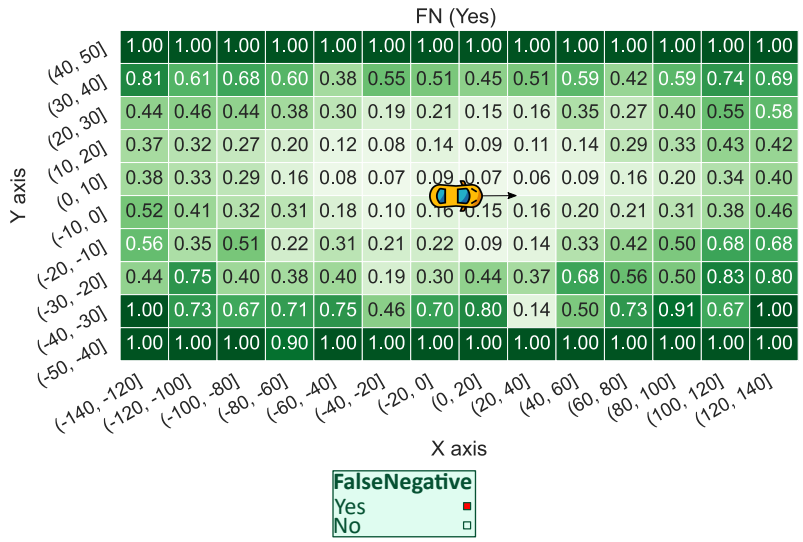


FIGURE 7.1: Performance limitation map for FN in the described scene and available data used for learning. Better performance of LIDAR is be observed near the Highly Automated Driving (HAD) vehicle.

- Better detection capability in the vicinity of the HAD vehicle is observed.
- FN is symmetrically distributed across X and Y axes with slightly higher FN distribution in front and on the right side of the HAD vehicle.

The uncertainty of the scene can be represented with a PLM (Fig. 7.1), which can be expressed as a quantitative evaluation of SOTIF for a given scene and system under consideration.

7.1.1.2 SOTIF Improvement Measures

The representation of PLM in the form of distance-based grid map (Fig. 7.1) provides the following SOTIF improvement measure.

- **System Modification**
  - LIDAR perception system with increased spatial range, if it is observed in the PLM that the FN distribution values for grid cells father from the vehicle is high and does not fulfill the design requirements.
  - LIDAR perception system with decreased FN, if it is observed in the PLM that the FN distribution values does not fulfill the design requirements.

7.1.1.3 Refinement

No localised refinement is provided at this stage based on the PLM.

7.1.2 Occlusion→FN

7.1.2.1 CPLM

CPLM related to $P(FN|Occlusion)$ is represented through Fig. 7.2,7.3,7.4. $P(FN = Yes|Occlusion = Fully\ visible)$ is represented by Fig. 7.2 and $P(FN = Yes|Occlusion = Largely\ Occluded)$ is represented by Fig. 7.3. Evidently $occlusion = largely\ occluded$

conditioned scenes have higher probabilities of FNs than $occlusion = fully\ visible$ scenes. The following analysis conclusions can be drawn.

- Fully visible scenes have considerably lower FN probability than fully visible scenes for LIDAR, given the dataset.
- The average $P(FN|Occlusion)$ probability is symmetrically distributed across X and Y axes with slightly higher FN probability in front and on the right side of the HAD vehicle.
- The *P(FN=Yes | Occlusion=Unknown)* results (Fig. 7.4) indicate the very low occurrence of the events. The values recorded for each cell are also abrupt and require further detail.

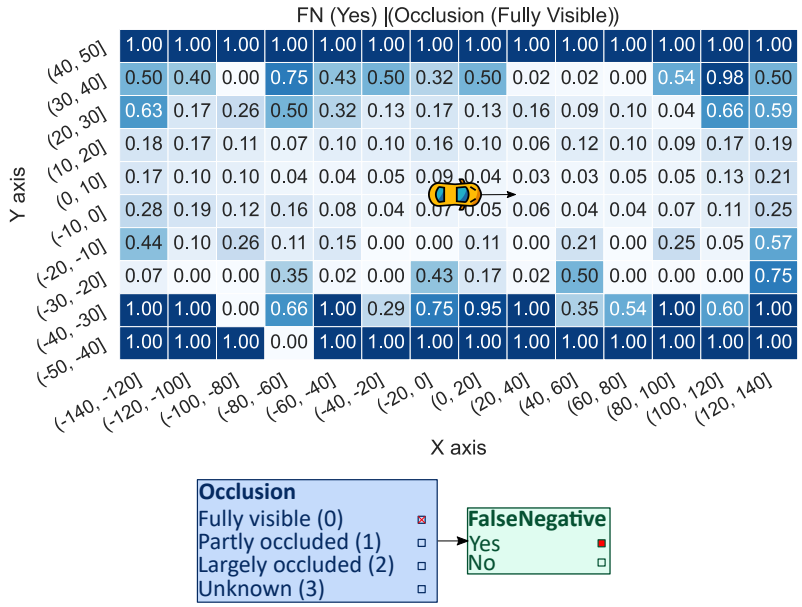


FIGURE 7.2: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on Occlusion (fully visible) in the described scene. CPLM for FN (yes) and occlusion (fully visible) scenes describe a low occurrence of FN in scenes that are labelled as fully visible.

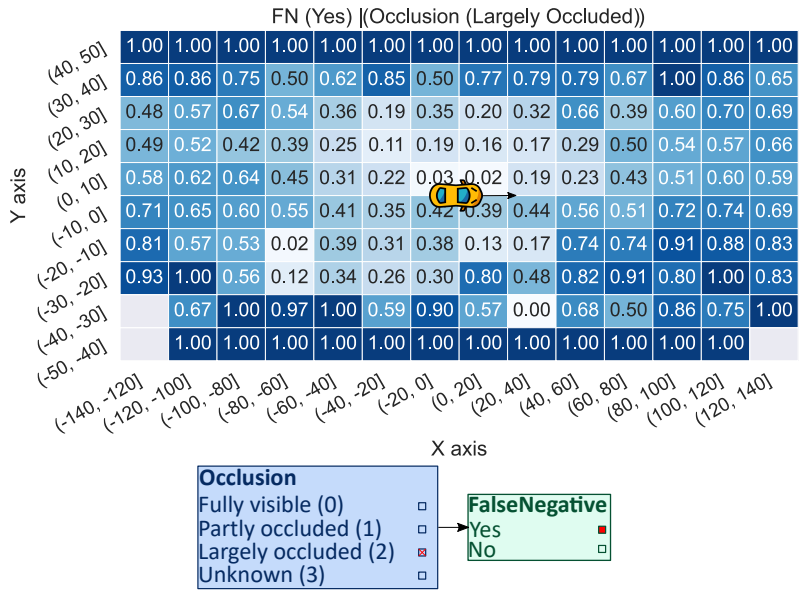


FIGURE 7.3: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on Occlusion (largely occluded) in the described scene. CPLM for FN (yes) and occlusion (largely occluded) scenes describe a high occurrence of FN in scenes that are labelled as largely occluded. Empty cells represent that no data instances were available.

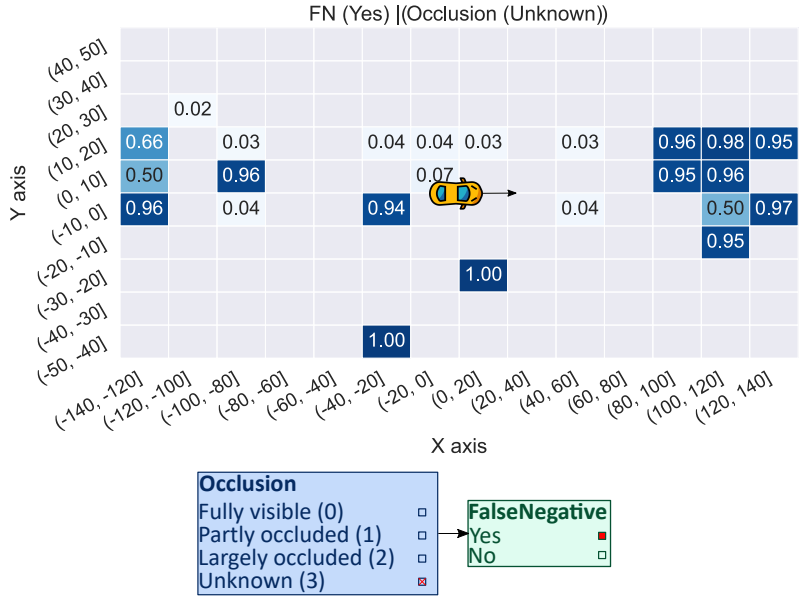


FIGURE 7.4: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on Occlusion (unknown) in the described scene. CPLM for FN (yes) and occlusion (unknown) scenes describe a low occurrence of data. Empty cells represent that no data instances were available.

7.1.2.2 Explicate Confidence

The results from implementation of PPMI are provided in a grid map format in Fig. 7.5 and 7.6. PPMI (Fig. 7.5,7.6) reflects confidence in the results. It is evident that events *FN=Yes* and *Occlusion=Largely Occluded* are dependent in comparison to

FN=Yes and *Occlusion=Fully Visible*, as per the PPMI variation interval defined in Sec 5.2.5.3. The experts deduce the following result.

- The confidence map in general supports the occlusion analysis conclusion (CPLMs).

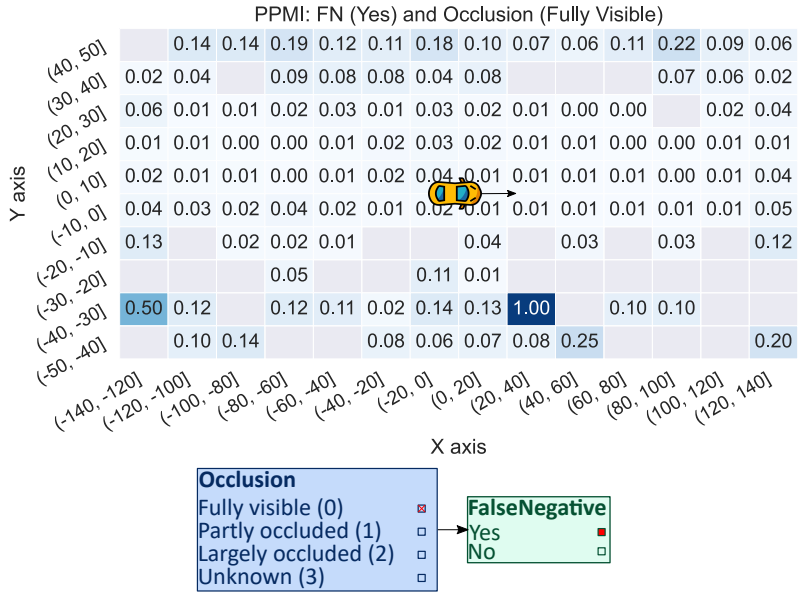


FIGURE 7.5: Positive pointwise mutual information representing the confidence measure on the FN (yes) and Occlusion (full visible). As the values are near 0, independence of states can be concluded. Empty cells represent that no data instances were available.

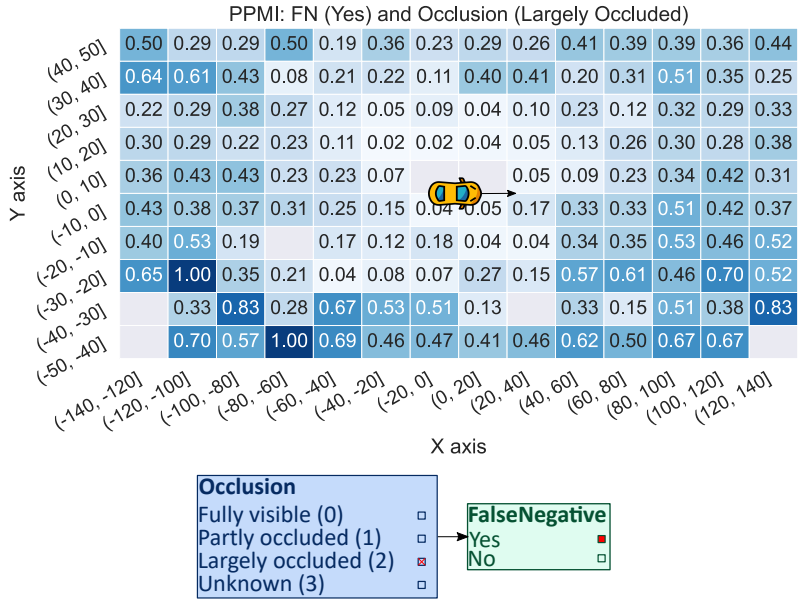


FIGURE 7.6: Positive pointwise mutual information representing the confidence measure on the FN (yes) and Occlusion (largely occluded). As the values are relatively higher than 0, some dependence between states can be concluded. Empty cells represent that no data instances were available. Empty cells represent that no data instances were available.

7.1.2.3 Localized Refinement

The following refinement steps can be taken.

- Refinement of causal factors for occlusion occurrence. This type of refinement is studied in the semi-automated discovery algorithm.
- Some regions in the grid map of $occlusion = fully\ visible$ have zero values surrounded by higher values. This abruptness in cell requires further data instance and analysis for robust results.
- In case of *P(FN=Yes | Occlusion=Unknown)*, further analysis of the corresponding scene is required by the expert.

7.1.2.4 SOTIF Improvement Measures

The following SOTIF improvement measures are proposed.

- **System Modification**
  - Inclusion of occlusion detection algorithm [83]. Such implementation and improvement will assist in prediction of FN probability at real time, thus predicting the performance of the SOTIF related HAD function.
- **Functional Restriction**
  - Restriction of HAD function in highly occluded scenes.

The system modification precedes functional restriction stated above. It results in decisions related to functional restrictions based on the identification of occluded scenarios.

7.1.3 Reflection→FN

7.1.3.1 CPLM

CPLMs related to *P(FN | Reflection)* are represented through Fig. 7.7 and 7.8. From the CPLMs, the following analysis conclusions can be drawn.

- Scenes with reflections have slightly higher cell values compared to no reflection scenes for LIDAR, given the dataset. However, the difference is relatively very small.
- Empty cells occur in the *P(FN | Reflection)* scenes especially in y directions indicating lesser dataset occurrences.

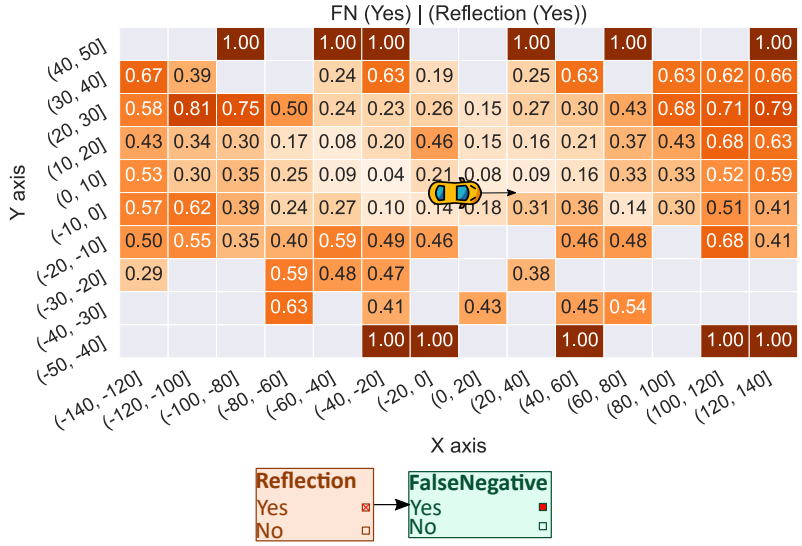


FIGURE 7.7: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on reflection (yes) in the described scene. CPLM for FN (yes) and reflection (yes) scenes describe a slightly higher occurrence of FN in scenes with reflections. Empty cells represent that no data instances were available.

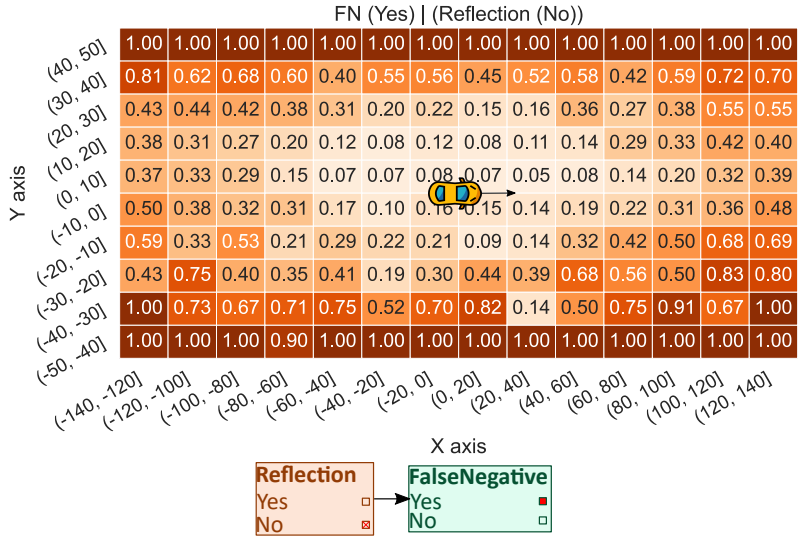


FIGURE 7.8: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on reflection (no) in the described scene. CPLM for FN (yes) and reflection (no) scenes describe a slightly lower occurrence of FN in scenes with reflections.

7.1.3.2 Explicate Confidence

The results from implementation of PPMI are provided in a grid map format in Fig. 7.9 and 7.10. It can be noticed that *FN=Yes* and *Reflection=Yes* are relatively less co-dependent than *FN=Yes* and *Reflection=No*. The experts deduce the following result.

- The confidence map does not support the reflection analysis conclusion (CPLMs).

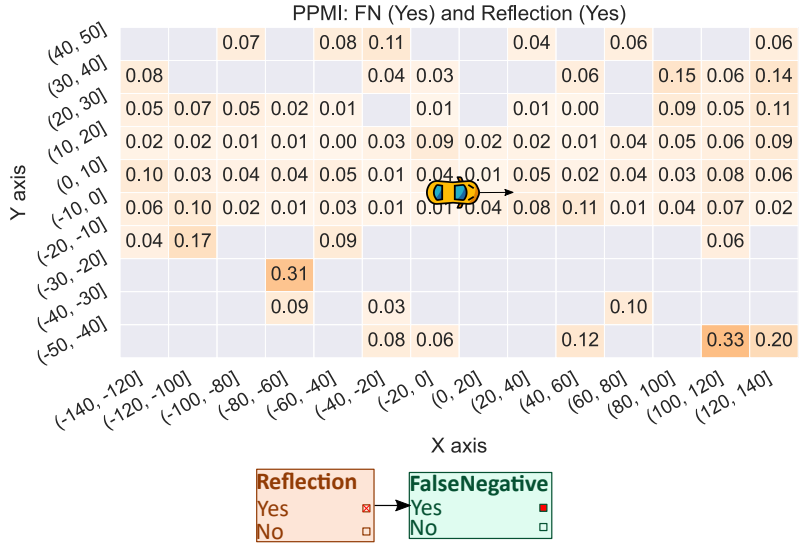


FIGURE 7.9: Positive pointwise mutual information representing the confidence measure on the FN (yes) and reflection (yes). As the values are towards 0, independence of states can be concluded. Empty cells represent that no data instances were available.

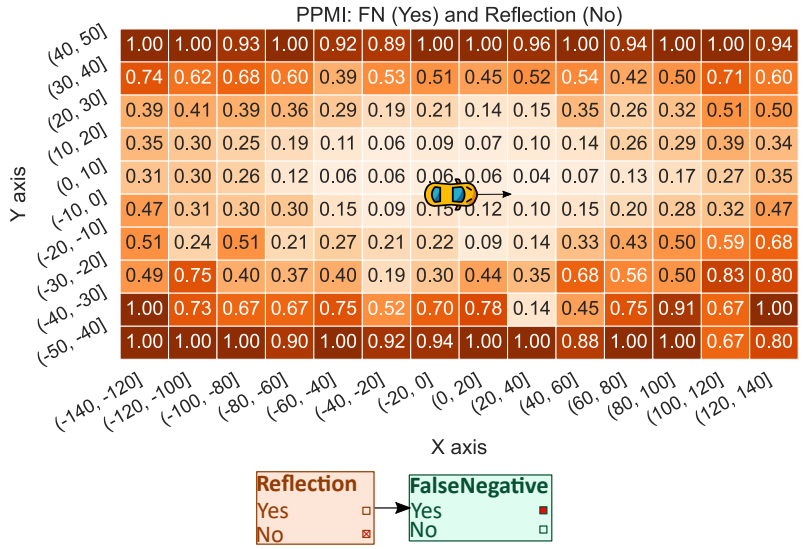


FIGURE 7.10: Positive pointwise mutual information representing the confidence measure on the FN (yes) and reflection (no). As the values are relatively higher than 0, some dependence between states can be concluded.

7.1.3.3 Localized Refinement

The following refinement step can be taken.

- Some regions in the grid map of $reflection = yes$ have no values available for cells. This requires further data collection and analysis for robust results.
- Reflection in general effect FP distribution of LIDAR detection which is also discussed in the literature [74]. Therefore, the experts propose further data collection and analysis of causal relations.

7.1.3.4 SOTIF Improvement Measure

No SOTIF improvement measure is proposed at this iteration of the analysis.

7.1.4 Truncation→FN

7.1.4.1 CPLM

CPLMs related to *P(FN | Truncation)* are represented through Fig. 7.12 and 7.11. From the CPLMs, the following analysis conclusions can be drawn.

- Low occurrences of *P(FN=Yes | Truncation=Yes)* are observed.
- The results indicate that truncation only causes FN at around relatively small radius of the HAD vehicle.

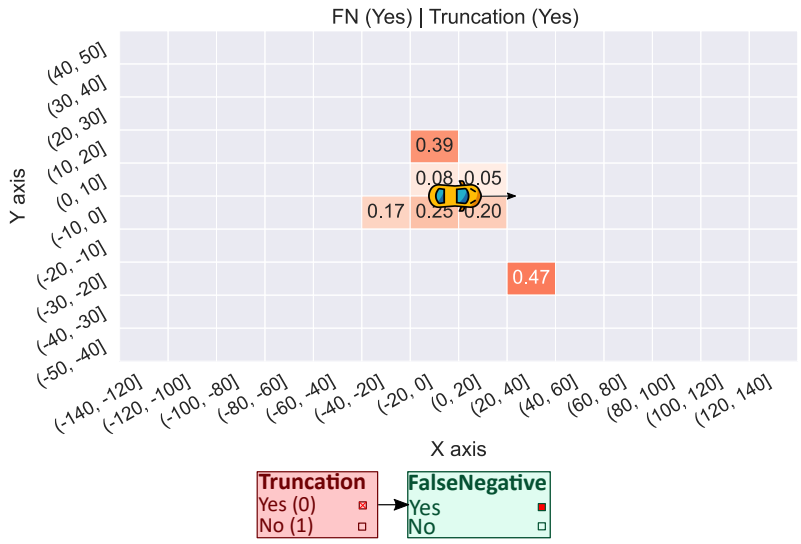


FIGURE 7.11: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on truncation (yes) in the described scene. CPLM for FN (yes) and truncation (yes) scenes have very low data representation. The result also indicates that truncation as a phenomenon occurs only in the vicinity of the Highly Automated Driving (HAD) vehicle. Empty cells represent that no data instances were available.

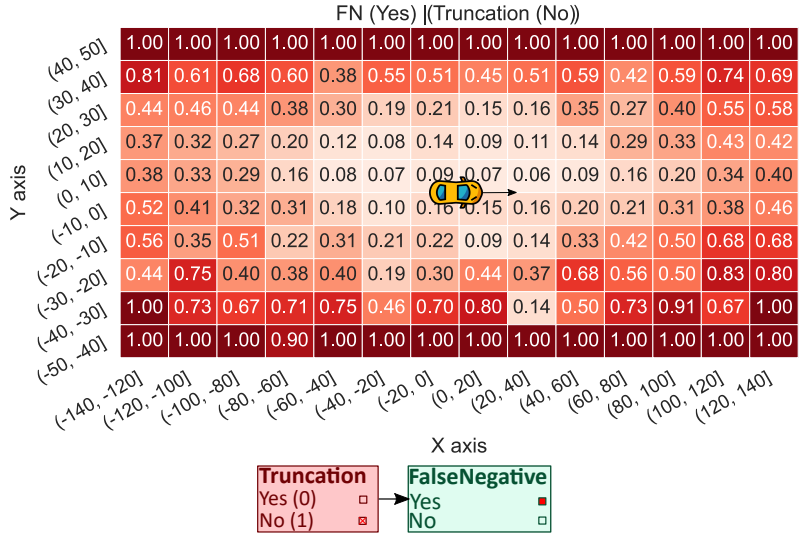


FIGURE 7.12: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on truncation (no) in the described scene.

7.1.4.2 Explicate Confidence

The results from implementation of PPMI are provided in a grid map format in Fig. 7.14 and 7.13. The experts deduce the following results.

- *FN=Yes and truncation=Yes* scenes' co-occurrence is negligible, further increasing the confidence about the more data required to understand the causal relation.

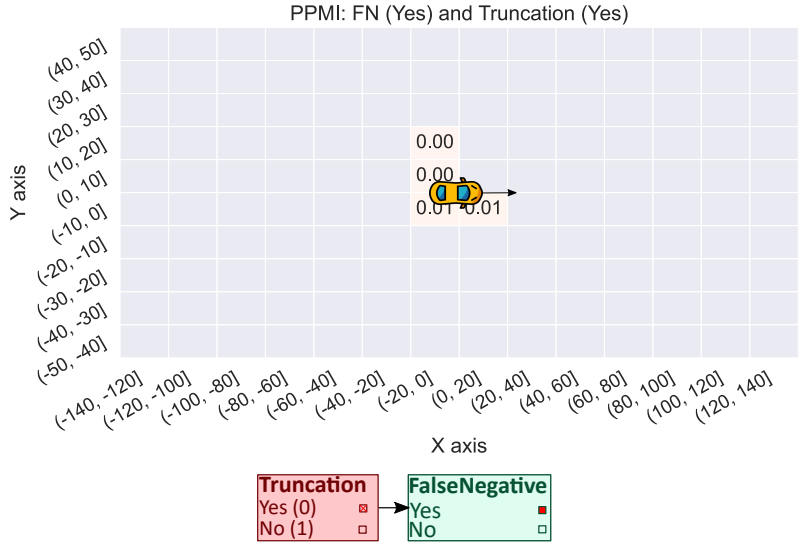


FIGURE 7.13: Positive pointwise mutual information representing the confidence measure on the FN (yes) and truncation (yes). Scarce data instances exist for this result. Empty cells represent that no data instances were available.

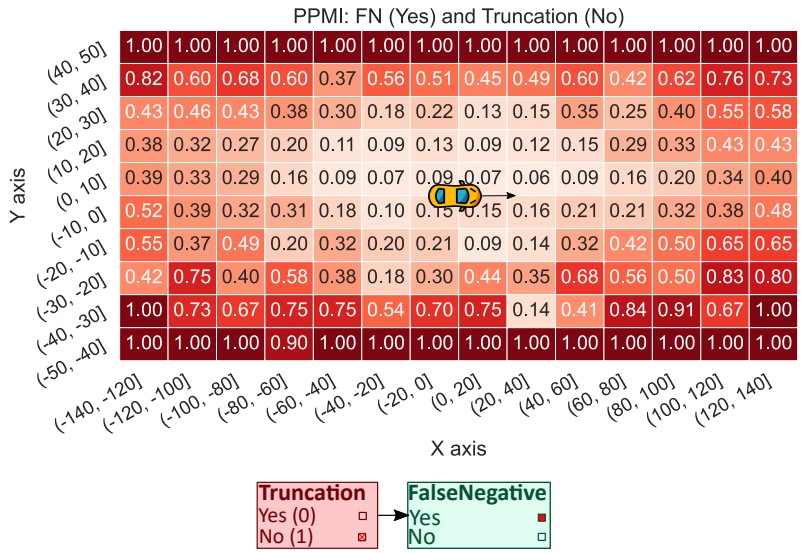


FIGURE 7.14: Positive pointwise mutual information representing the confidence measure on the FN (yes) and truncation (no).

7.1.4.3 Localized Refinement

The following refinement step can be taken.

- More data is required for truncation node in order to establish or negate a causal relation.

7.1.4.4 SOTIF Improvement Measures

No SOTIF improvement measure is proposed at this stage of the analysis.

7.1.5 Weather→FN

7.1.5.1 CPLM

CPLMs related to *P(FN | Weather)* are represented through Fig. 7.15 and 7.16. From the CPLMs, the following analysis conclusion can be drawn.

- Grid map representing rainy weather scenes have slightly higher cell values than sunny weather scenes.

Causal relation between FN and rain for LIDAR detection is well documented in literature [45]. A relatively slight increase in the $P(FN = yes|weather = rain)$ than $P(FN = yes|weather = sunny)$ can be attributed to light rain phenomena presence in the used dataset.

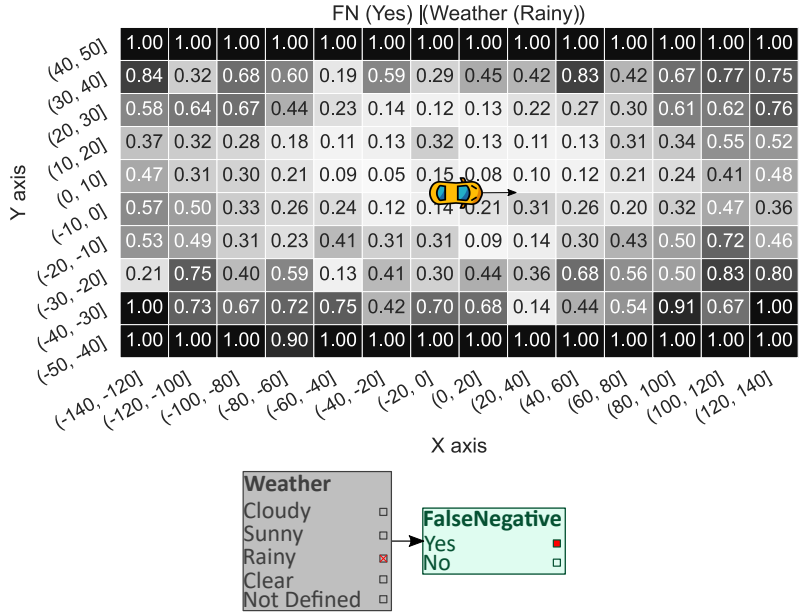


FIGURE 7.15: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on weather (rain) in the described scene. CPLM for FN (yes) and weather (rain) scenes describe a slightly higher occurrence of FN values.

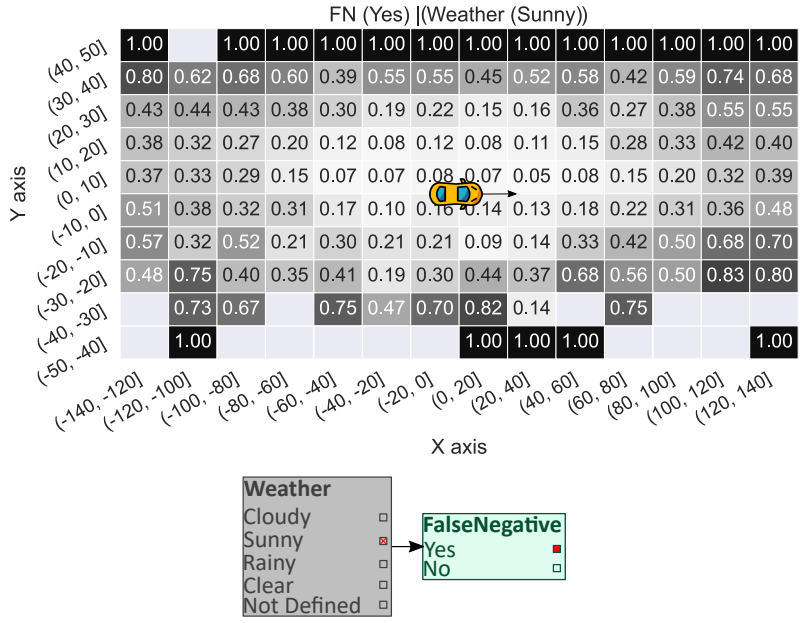


FIGURE 7.16: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on weather (sunny) in the described scene. CPLM for FN (yes) and weather (sunny) scenes describe a slightly lower occurrence of FN values. Empty cells represent that no data instances were available.

7.1.5.2 Explicate Confidence

The results from implementation of PPMI are provided in a grid map format in Fig. 7.17 and 7.18. It can be noticed that *FN=Yes* and *Weather=Rainy* are relative more co-dependent than *FN=Yes* and *Weather=Sunny*. However, at *Y=[30,40]* in Fig. 7.18

some anomalously high values can be observed. The expert deduces the following result.

- The confidence map in general supports the weather analysis conclusion (CPLMs).

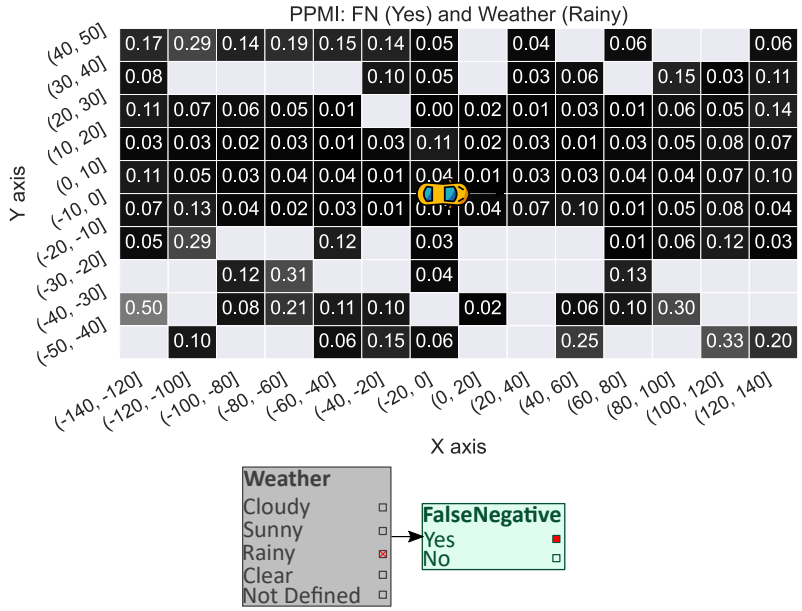


FIGURE 7.17: Positive pointwise mutual information representing the confidence measure on the FN (yes) and weather (rainy). Empty cells represent that no data instances were available.

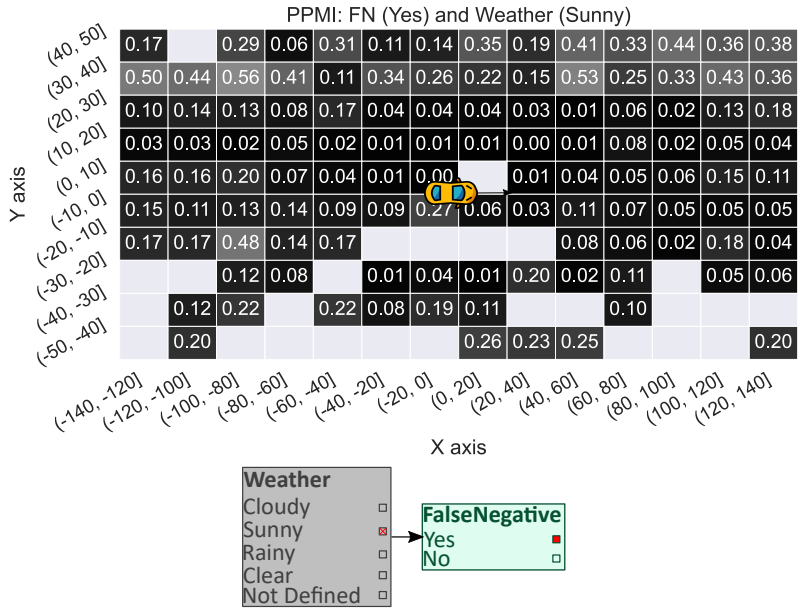


FIGURE 7.18: Positive pointwise mutual information representing the confidence measure on the FN (yes) and weather (sunny). Empty cells represent that no data instances were available.

7.1.5.3 Localized Refinement

The following refinement step can be taken.

- In order to further plausibilize the causal effect of rain on the FN, further data instances which include heavy rain phenomena are required.

7.1.5.4 SOTIF Improvement Measures

The following SOTIF improvement measures are proposed.

- **System Modification**
  - Inclusion of weather sensor: Such improvement will assist in prediction of FN probability at real time, thus predicting the performance of the SOTIF related HAD function.
  - Modification of the fusion algorithm, if applicable.
- **Functional Restriction**
  - Restriction of HAD function in heavy rain.

The system modification precedes functional restriction stated above. The system modification results in decisions related to functional restrictions based on the identification of rainy weather scenes.

7.2 Knowledge Discovery Algorithm Results

This section provides results representative of validating, refine and augment block (Fig. 5.4) and the algorithm **knowledge discovery**. The block supports the identification of anomalous and rarely occurring scenes which are then analysed for potential novel triggering conditions. For simplicity and completeness, the following patterns are adopted in the result.

- Hypothesis testing and relevant scene identification is provided for *FN*, *truncation*, *occlusion* and *reflection* nodes from the initial CBN (Fig. 7.19).
- *Traffic density*, *vehicle activity* and *context* are selected from the relevant scene causal relation step for refinement and validation (Fig. 7.20(a)). Moreover, the refinement step is performed for all three selected casual relations with respect to *FN* only.
- CCET refinement and corresponding validation is performed on *traffic density* with *FN* and its two parent nodes i.e., *reflection* and *occlusion* (Fig. 7.20(b)).
- Triggering condition removal step is performed on the *truncation* node (Fig. 7.20(a)).
- The updated CBN structure is again tested through estimate and plausibilize block (knowledge plausibilization algorithm).

7.2.1 Relevant Scene Identification

Random division of train and test datasets results in varying number of relevant scenes' identification. Number of relevant scenes identified also vary based on the relevant nodes for which the relevant scene identification is performed. The $RSS^{node}_{initial}$ for *node:* FN, occlusion and reflection is shown in Fig. 7.19. The number of scenes identified for all the relevant nodes decrease generally (3.68% of the test dataset scenes on average) from the total number of scenes in the test dataset. A relatively

small number of scenes identified for a specific node (e.g., for reflection in Fig. 7.19) that the train and test datasets are similar for the specific node under consideration.

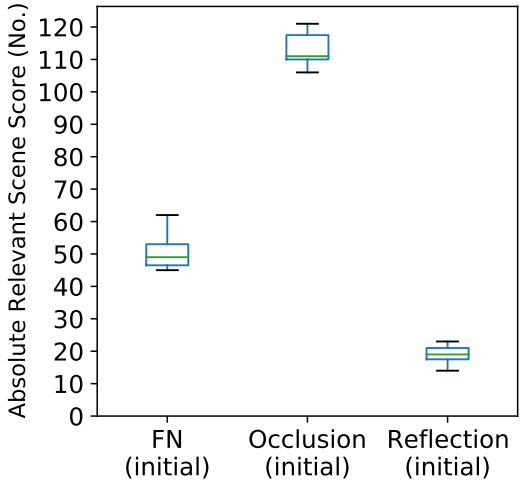


FIGURE 7.19: Relevant Scene Score (RSS) for hypothesis testing of FN, occlusion and reflection. RSS here represents the numbers of scenes rarer at significance level α before any modification in the CBN structure.

7.2.2 Relevant Scene Causal Relations

In order to identify relevant scene causal relations, it is assumed that the RSS provides the best scope. Based on the assumption, the variables with the highest scene score are considered for further analysis.

TABLE 7.1: Relevant scene causal relations and the respective acceptable random occurrences and identifiable novel triggering conditions cases. Identification of the triggering conditions is based on the expert opinion.

Relevant Scene Causal Relations	FN	Occlusion	Reflection
Acceptable Random Occurrences	29	56	10
Novel Triggering Condition Identifiable	33	65	13

7.2.2.1 Acceptable Random Occurrences

Out of the identified relevant scenes from FN, occlusion and reflection test, the expert analysis concludes in 29, 56 and 10 scenes as acceptable random occurrences. This indicates that no potential novel triggering conditions can be identified in those scenes (Tab. 7.1). This conclusion does not necessarily mean that no novel triggering condition is present in those scenes. The conclusion supports the fact that the experts cannot provide a probable explanation of the scene relevancy from the novel triggering condition viewpoint.

7.2.2.2 Novel Triggering Condition Identifiable

Out of the identified relevant scenes from FN, occlusion and reflection test, the expert analysis concludes in 35, 65 and 13 scenes as representatives of some potential novel triggering conditions (Tab. 7.1). The potential novel triggering conditions identified for FN hypothesis testing are listed in Tab. 7.2. Identification of potential novel triggering conditions is performed through analysis of individual scene e.g., Fig. 5.8 represents *"ground truth labelling errors"*, *"vehicle activity"* and *"cars loaded on a trailer"* as the expert opinion about the potential novel triggering conditions present in the scene.

TABLE 7.2: Potential novel triggering conditions identified by FN's p-value hypothesis testing. Expert analyses the relevant scenes identified through hypothesis testing. Triggering conditions are then provided by the experts.

Potential Novel Triggering Conditions	No. of Occurrences
Traffic Density	20
Vehicle Dimensions	20
Cars loaded on a trailer	5
Ground Truth Labelling Error	8
Lane Discretization	13
Construction	2
Divider on Road	1
Other lane height	3
Context	7
Vehicle Activity	11

7.2.3 Refinement

In the refinement step, only *traffic density*, *context* and *vehicle activity* out of the potentially novel triggering conditions mentioned in Tab. 7.2 are discussed. The selection is made for the following reasons.

- Relatively high number of occurrences is observed for the selected causal relations (Tab. 7.2).
- Labelled data is available for the selected causal relations. The decision is specific to the scope of implementation in this thesis. In other instances, labelled data must be made available for any selected causal relation.

The experts provide the following propositions.

- *Traffic density*, *vehicle activity* and *context* may affect the selected performance limitation i.e., *FN* (Fig.5.9(a)).
- *Traffic density* may affect *FN*, *occlusion* and *reflection*. This proposition specifically focuses on the CCET case (Fig.5.9(b)).
- *Truncation* may not have any effect on the defined performance limitation i.e., *FN*.

As discussed in Sec. 5.3.2.8 CCET can be established through DCET by combining two individual DCET of parent and child node as well as the expert opinion. Both DCET and CCET are implemented only for novel triggering condition, traffic density (Sec. 5.3.2.8). Expert's intuition about the causal relation can support the decision of DCET or CCET selection e.g., apart from the SOTIF related performance limitation (FN), the context of driving and the vehicle activity does not support the intuition of the cause or effect of any other node in the CBN.

7.2.4 Validation

The expert analysis of the results by the implementation of Algorithm 1 provides the validation.

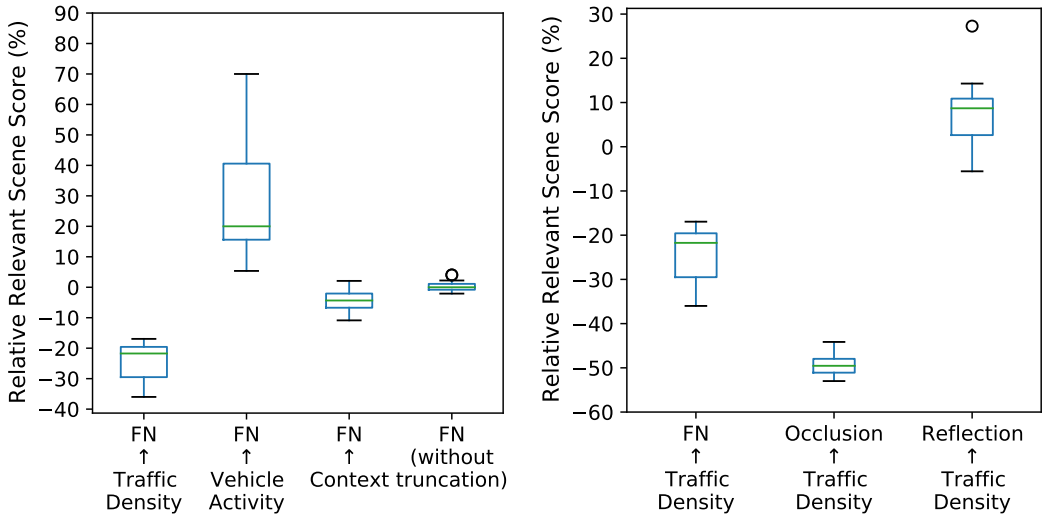


FIGURE 7.20: **(a)** Relative relevant scene score of refinement of Direct Causal Edge Trail (DCET) for FN using traffic density, vehicle activity, context and without truncation as identified novel triggering conditions. Lower value after the DCET refinement indicates a more suitable Causal Bayesian Network (CBN) structure than the structure considered before. **(b)** Relative relevant scene score of refinement of traffic density as DCET of FN, occlusion and reflection. The circles in the plot refer to the outlier relative Relevant Scene Score (RSS) among the iterations.

7.2.4.1 Traffic Density

The validation results of the DCET (Sec. 5.3.2.8) performed for the traffic density are shown in Fig. 7.20(a) and Fig. 7.20(b). The validation results that can be extracted by implementing Algorithm 1 are following.

$$RSS^{FN}_{initial} > RSS^{FN}_{after} \implies proposition_{Traffic\ Density \rightarrow FN} = valid \tag{7.1}$$

$$RSS^{Occ}_{initial} > RSS^{Occ}_{after} \implies proposition_{Traffic\ Density \rightarrow Occlusion} = valid \tag{7.2}$$

$$RSS^{Ref}_{initial} < RSS^{Ref}_{after} \implies proposition_{Traffic\ Density \rightarrow Reflection} = invalid \tag{7.3}$$

The experts draw the following conclusions based on the proposition equations (Eq. 7.1,7.2,7.3).

- Traffic density may have a causal effect on FN and occlusion.
- Traffic density causal relation with the FN and occlusion may be formally represented with a CCET case.
- Traffic density may not be a causal factor for reflection.

7.2.4.2 Context

The validation results of the DCET (Sec. 5.3.2.8) performed for the context of driving as triggering condition to FN is shown in Fig. 7.20(a).

The validation result that can be extracted by implementing Algorithm 1 is following.

$$RSS^{FN}_{initial} > RSS^{FN}_{after} \implies proposition_{Context \rightarrow FN} = valid \tag{7.4}$$

A general decrease in the relevant scenes after the adjustment in the CBN structure is observed. However, the decrease in the RSS is not substantial (-4.20%), apparently due to skewed data representation (199645 data points representing highway and 863 data points representing urban) of the highway and urban context, from the experts' standpoint. The experts draw the following conclusion based on the proposition equations (Eq. 7.4).

- Context may have a causal effect on the FN; however, more data is required to refute or substantiate this claim.

7.2.4.3 Vehicle Activity

The validation results of the DCET (Sec. 5.3.2.8) performed for the vehicle activity as triggering condition to FN is shown in Fig. 7.20(a). The validation result that can be extracted by implementing the Algorithm 1 is following.

$$RSS^{FN}_{initial} < RSS^{FN}_{after} \implies proposition_{Vehicle\ Activity \rightarrow FN} = invalid \tag{7.5}$$

An increase is observed in the RSS after the adjustment in the CBN structure (29.02%). The experts draw the following conclusion based on the proposition equation (Eq. 7.5).

- The proposition that vehicle activity is a triggering condition for FN is not valid, given the datasets. Thus, vehicle activity cannot be taken as a triggering condition for FN at this point.

7.2.4.4 Truncation Removal

The validation results of the DCET (Sec. 5.3.2.8) performed for the truncation removal as triggering condition to FN is shown in Fig. 7.20(a). The validation result that can be extracted by implementing the Algorithm 1 is following.

$$RSS^{FN}_{initial} < RSS^{FN}_{after} \implies proposition_{Truncation \rightarrow FN} = valid \tag{7.6}$$

A slight increase is observed in the RSS when truncation is not taken as a triggering condition for FN. Algorithm 1 provides the validity of the proposition. The experts draw the following conclusion based on the proposition equation (Eq. 7.6).

- Truncation removal proposition has inconclusive evidence. Further collection of data is required to refute or substantiate further claim.

Tab. 7.3 summarizes the results of the implementation. The results validated and analysed by the experts to provide final conclusions using Algorithm 1.

TABLE 7.3: Summary of the results produced by the implementation of the methodology. Different initial nodes, systematic factors and refinements are considered in the implementation.

Initial CBN Node	**Relative RSS** (%)	**Novel Triggering Condition**	$proposition_{NTC}$ **(Algorithm. 1)**	**Expert Conclusion (Algorithm. 1)**
FN	29.02	Vehicle Activity	Invalid	Rejected Proposition
FN	-4.20	Context	Valid	Inconclusive Evidence
FN	-24.52	Traffic Density	Valid	Accepted Proposition
Occlusion	-49.50	Traffic Density	Valid	Accepted Proposition
Reflection	7.82	Traffic Density	Invalid	Rejected Proposition
FN	0.45	*Without Truncation*	Invalid	Inconclusive Evidence

The augmented CBN model that results from the implementation of semi-automated discovery of triggering conditions is shown in Fig. 7.21. Traffic density is adjusted as CCET for FN and occlusion, representing a confounding phenomenon. Similarly, further data collection can assist in the analysis and decision-making about truncation and driving context as part of the triggering conditions. In this manner, novel triggering conditions identification can be performed, refinements can be generated and validated through an iterative process to acquire more knowledge and perform a robust SOTIF analysis.

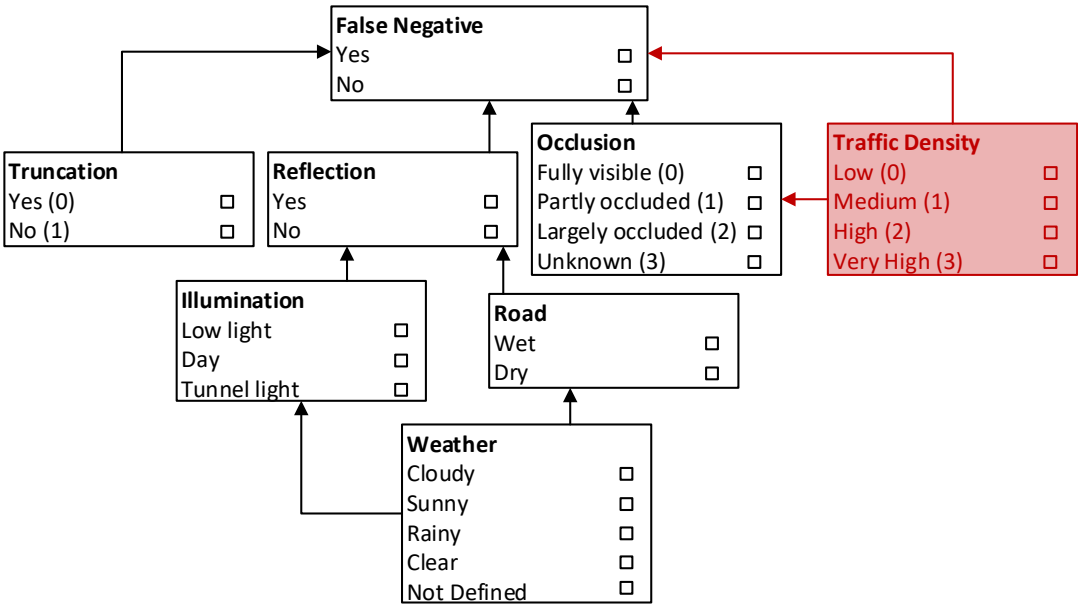


FIGURE 7.21: CBN structure from Fig. 6.1 updated with novel triggering condition *traffic density* as SOTIF relevant scenario factor for false negative and occlusion.

7.3 Knowledge Plausibilization Algorithm Results: Second Iteration

As the CBN structure is updated, new results should be introduced from the estimate and plausibilize block iteration. At this stage, the following results need to be added and amended.

- Inclusion of traffic density requires the analysis of the node.
- The confounding phenomena of traffic density (Fig. 7.21) requires reassessment of the causal relation between FN and occlusion.

7.3.1 Occlusion→FN

7.3.1.1 CPLM

CPLM related to $P(FN|do(Occlusion))$ is represented through Fig. 7.22, Fig. 7.23 and Fig. 7.24. Intervention results instead of associational ones (Sec. 5.2.4.2) are necessary at this stage as traffic density effect on the causal relation needs to be marginalized [87]. This exercise becomes important in cases where some threshold as an acceptable criterion is defined. The following analysis conclusion can be drawn.

- Fully visible scenes have considerably lower FN probability than largely occluded scenes for LIDAR, given the dataset.
- The average $P(FN|do(Occlusion))$ probability is symmetrically distributed across X and Y axes with slightly higher FN probability in front and on the right side of the HAD vehicle for fully visible scenes.
- The average $P(FN|do(Occlusion))$ probability is symmetrically distributed across X and Y axes with considerably higher FN probability in front and on the right side of the HAD vehicle for largely occluded scenes.

- The *P(FN=Yes | do (Occlusion=Unknown))* results (Fig. 7.24) indicate the very low occurrence of the events. Less abruptness in the values of the cells is observed in the interventional results (Fig. 7.24) than the associational results (Fig. 7.4). One reasoning can be attributed to the normalization of confounding effect on the traffic density by measuring intervention.

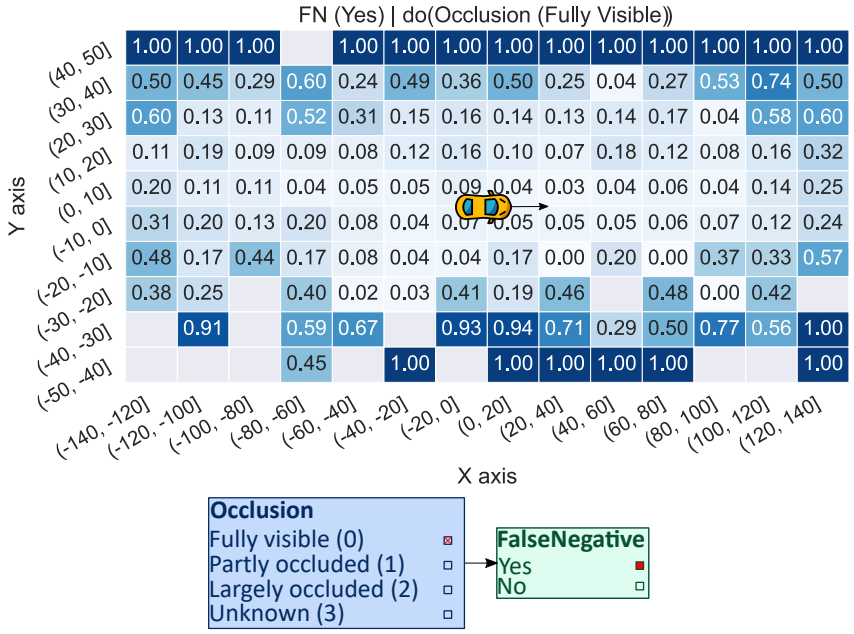


FIGURE 7.22: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on Occlusion (fully visible) in the described scene. CPLM for FN (yes) and occlusion (fully visible) scenes describe a low occurrence of FN in scenes that are labelled as fully visible. The interventional quantity marginalizes the effects of confounding variables and provides the causal relations. Empty cells represent that no data instances were available.

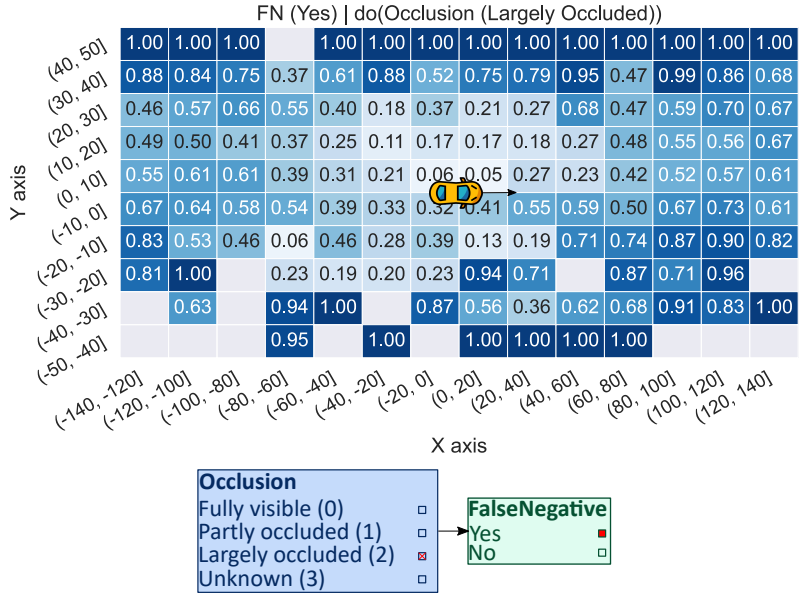


FIGURE 7.23: Conditional Performance Limitation Map (CPLM) for FN (Yes) conditioned on Occlusion (largely occluded) in the described scene. CPLM for FN (Yes) and occlusion (largely occluded) scenes describe a high occurrence of FN in scenes that are labelled as largely occluded. The interventional quantity marginalizes the effects of confounding variables and provides the causal relations. Empty cells represent that no data instances were available.

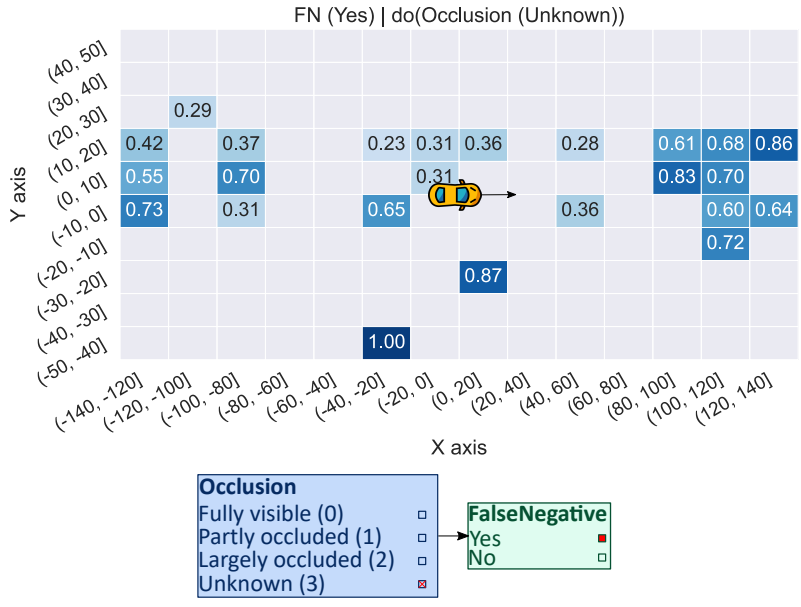


FIGURE 7.24: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on Occlusion (unknown) in the described scene. CPLM for FN (yes) and occlusion (unknown) scenes describe a low occurrence of data. The interventional quantity marginalizes the effects of confounding variables and provides the causal relations. Empty cells represent that no data instances were available.

The importance of *do*-operator and interventional calculation is evident. It marginalizes the effects of confounding variables providing better results. This adjustment

becomes necessary for safety when decisions are made on the probability values. As an example, consider the cell $X = (40, 60)$ and $Y = (30, 40)$ of Fig. 7.3 and Fig. 7.23. Fig. 7.3 represents the associational result $(P(FN = yes|Occlusion = Largely\ Occluded))$ and Fig. 7.23 represents the interventional result $(P(FN = yes|do\ (Occlusion = Largely\ Occluded)))$. There is a significant difference between the two values 0.79 versus 0.95. If a threshold value at $\tau_2 = 0.8$ was chosen for this cell and no interventional result was acquired, SOTIF might have been wrongly accepted to be sufficient.

7.3.2 Traffic Density→FN

7.3.2.1 CPLM

CPLMs related to $P(FN|Traffic)$ are represented through Fig. 7.25, Fig. 7.26 and Fig. 7.27. The following analysis conclusion can be drawn.

- Low traffic density scenes have considerably lower FN probability than high traffic density scenes for LIDAR, given the dataset.
- Low and medium traffic densities scenes have considerably higher FN probability near the HAD vehicle than high traffic density scene for LIDAR, given the dataset. This can be attributed to the fact that in high density cases more static vehicles are present near the HAD vehicle and detection rate per scene increases consequently.

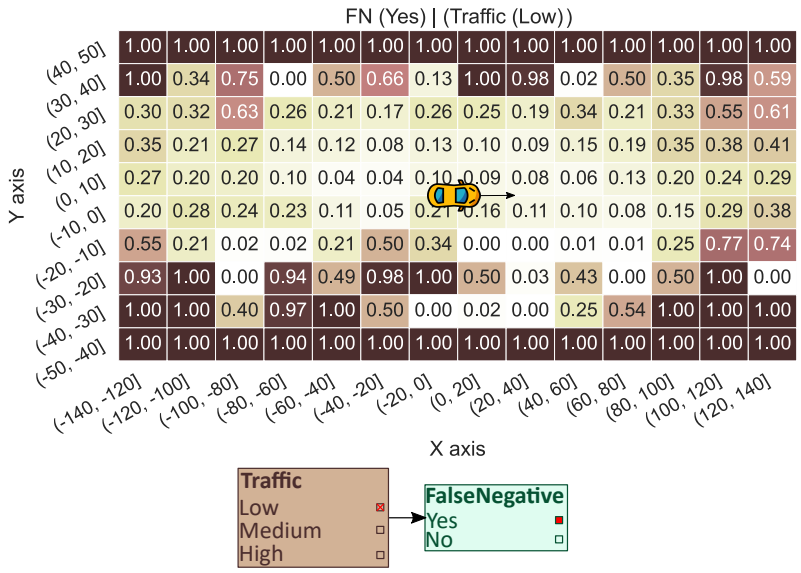


FIGURE 7.25: Conditional Performance Limitation Map (CPLM) for FN (Yes) conditioned on traffic (low) in the described scene. CPLM for FN (yes) and traffic (low) scenes describe a low occurrence of FN in scenes and locations farther from the Highly Automated Driving (HAD) vehicle that are labelled as traffic=low.

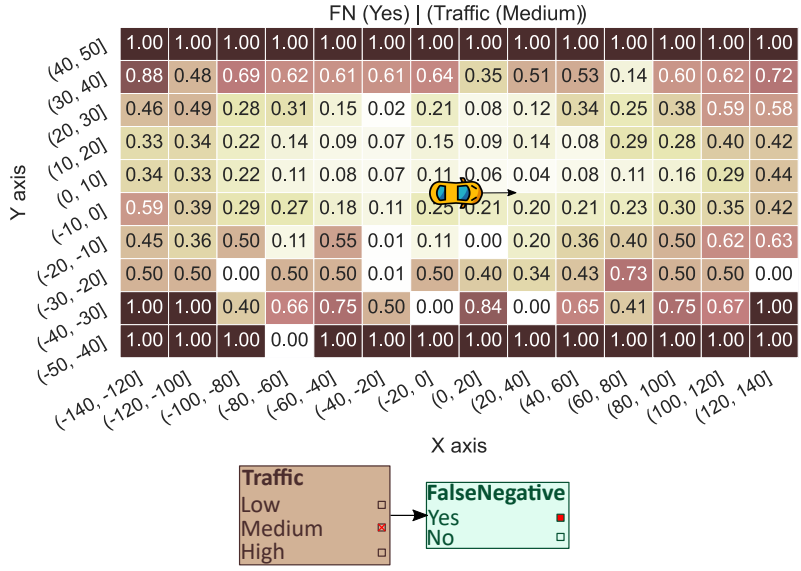


FIGURE 7.26: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on traffic (medium) in the described scene.

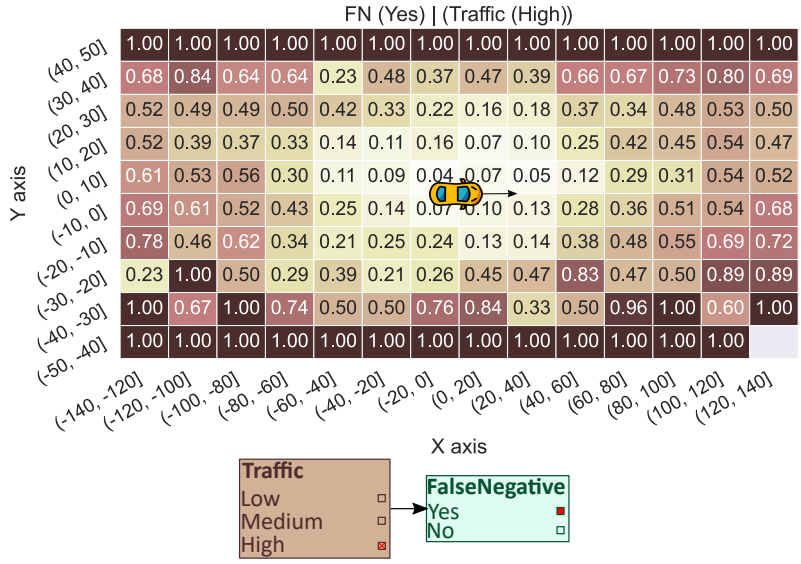


FIGURE 7.27: Conditional Performance Limitation Map (CPLM) for FN (yes) conditioned on traffic (high) in the described scene. CPLM for FN (yes) and traffic (high) scenes describe a high occurrence of FN in scenes and locations farther from the Highly Automated Driving (HAD) vehicle that are labelled as traffic=high.

7.3.2.2 Explicate Confidence

The results from implementation of PPMI are provided in a grid map format in Fig. 7.28, Fig. 7.29 and Fig. 7.30. It is evident that event *FN=Yes* and *Traffic=High* are co-dependent in comparison to *FN=Yes* and *Traffic=Low OR Medium*, as per the variation interval defined in Sec 5.2.5.3. The experts deduce the following results.

- The confidence map in general supports the $Traffic \rightarrow FN$ analysis conclusion.

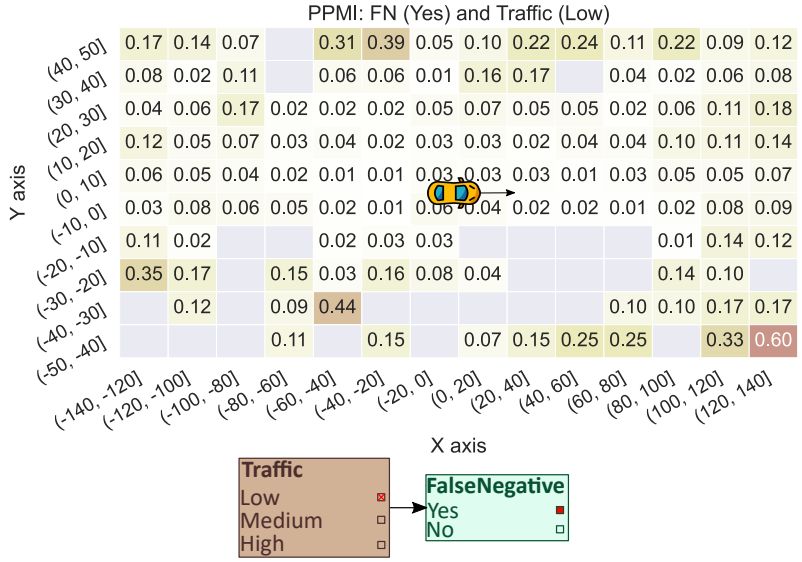


FIGURE 7.28: Positive pointwise mutual information representing the confidence measure on FN (yes) and traffic (low). As the values are only slightly above 0, very little dependence between states can be concluded. Empty cells represent that no data instances were available.

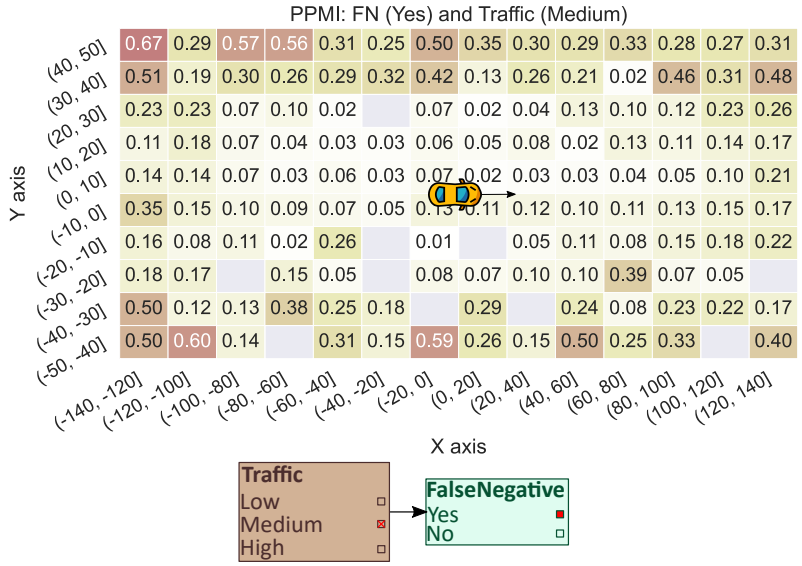


FIGURE 7.29: Positive pointwise mutual information representing the confidence measure on the FN (yes) and traffic (medium). As the values are only slightly above 0, very little dependence between states can be concluded. However, the values are greater than the case of traffic=low (Fig. 7.28). Empty cells represent that no data instances were available.

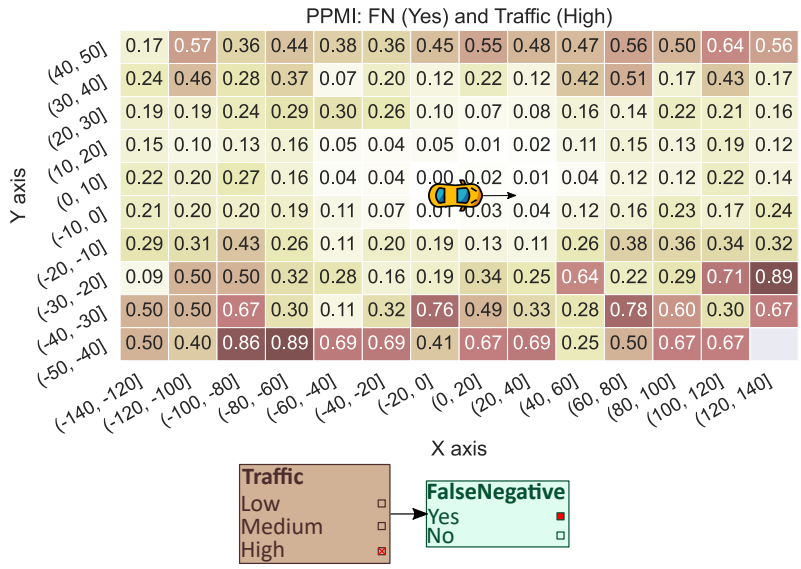


FIGURE 7.30: Positive pointwise mutual information representing the confidence measure on the FN (yes) and traffic (high). The values are relatively larger than 0 for cells farther from the Highly Automated Driving (HAD) vehicle, some level of dependence can be concluded between states.

7.3.2.3 Localized Refinement

The following refinement step can be taken.

- Refinement of causal factors for traffic density occurrence e.g., time of day, region etc. The results presented here support the refinement step of occlusion.

7.3.2.4 SOTIF Improvement Measures

The following SOTIF improvement measure is proposed.

- **System Modification**
  - Inclusion of traffic density detection algorithm [32]. Such implementation and improvement will assist in prediction of FN probability at real time, thus predicting the performance of the SOTIF related HAD function.

7.4 Summary

This chapter provides the implementation of the algorithms discussed in Ch. 5. Estimation and plausibilization of the initial CBN variables is provided first, followed by the implementation of the semi-automated discovery, which results in *traffic density* as novel triggering condition. A second iteration of estimation and plausibilization is provided to serve the newly discovered knowledge about the CBN.

It can be argued that the results provided support the contributions claimed in Ch. 5. Chapter 5 made the following contribution claims.

(S1) Measurement metrics and explanation of performance limitations

(S2) Relevant triggering conditions extraction

(S3) Convergence towards a manageable set of triggering conditions

(S4) Derivation of open context model from the SOTIF standpoint

(S5) Catalogue of abstract scenarios

(S6) Evaluation of the causal effect of one or more triggering conditions on performance limitations

(S7) Confidence building on the identified causal effect

(S8) Targeted SOTIF oriented modification of HAD function or ODD based on the causal effect calculations

(S9) An iterative SOTIF framework to assist semi automated discovery of triggering condition

The result section provides a PLM for FN *(S1)*. Relevant triggering conditions are extracted and discovered e.g., *traffic density (S2, S3, S9)*. The CBN supports an abstract definition of the open context *(S4)*. *Occlusion* and *traffic density* should be considered important for scenario catalogue formalization *(S5)*. The CBN provides calculations of the causal effect of multiple triggering conditions *(S6)*. PPMI metrics are used to provide confidence measures on the causal effect *(S7)*. Moreover, some available algorithms in literature are also considered as SOTIF improvement measures in order to manifest sufficient usability of the analysis *(S8)*.

In summary, it can be argued that the evaluation activities presented in this chapter provide a seamless implementation of the causal framework discussed in the last chapter. The evaluation supports the systematic modelling, identification and discovery of the perception performance limiting triggering conditions in automated driving. The conceptual framework presented is also well suited to implement for SOTIF analysis and supports Clause 7, 8, 10 and 11 of the standard [54].

8

Limitations and Threat to Validity

"Science never gives up searching for truth, since it never claims to have achieved it."

– John Charles Polanyi, *Hungarian Scientist*

The framework presented in this thesis poses limitation and threat to its validity at the theoretical, methodological and implementation level. In this chapter, some of the most important limitations of the framework and its implementation are discussed in detail, giving a complete picture of the introduced scientific method.

8.1 Theoretical Level Limitations

8.1.1 Data Abstraction and ODD Taxonomy

Every scene or scenario about the real world is represented at some level of abstraction e.g., data discretization of continuous variable in CBN [130]. Different abstraction may result in different maps e.g., a lower abstraction of *illumination* will be states with light intensities instead of *day* as state, a further lower abstraction for the same node will be taking a continuous light intensity distribution [1]. Different maps emanating from such selection will challenge the robustness of the results. It is believed that these abstractions can be governed by ODD taxonomies. Hence, a well-established ODD taxonomy can be used as the reference for data abstraction. Moreover, another possible solution can be dynamic discretization for optimization of robust results [35].

8.1.2 Argumentation on Completeness

The framework presented in this thesis provides a representative, robust and knowledge acquisition-based process. However, providing argumentation on the completeness of the safety model by our methodology to assess SOTIF is still a challenging task. The framework though also supports iterations for structure augmentation and data acquisition, the resulting CBN may still not be complete. A conventional solution can be expert decision and sufficiently small, rarer scenes (RSS) in the testable implication or high confidence on the plausibilized relation (e.g., PPMI maps), however the resulting CBN may still not be the best possible representation of the real world. Consider the following hypothetical scenario.

After the CBN restructuring with the traffic density including as a novel triggering condition, the RSS and expert conclude that a certain benchmark is satisfied and no further analysis on these nodes is required. However, it is possible that traffic density may not be confounding phenomenon and its effect on the FN is mediated

in by a third variable, assumed missing at this iteration. Such challenges can be partially addressed by understanding the underlying intuition of the causal relations about the real world.

8.1.3 Randomness and Lack of Knowledge Decoupling

Hypothesis testing provides a general conception of the acceptance of randomness in the results to a certain significance level α. Any value above or below the prescribed significance level α (depending on which tail of the distribution is being tested) is deemed as unacceptable randomness and results in the rejection of the hypothesis. In this work, instead of rejection of the hypothesis, it is modelled through identified novel potential triggering conditions. In its essence, this step corresponds to modelling lack of knowledge concepts [41].

The decoupling between randomness and lack of knowledge at some significance level α works with the underlying assumption (Assumption 9). However, this assumption can be violated and the rare scene occurrence can be purely governed by randomness in the dataset. This limitation is partially addressed by allowing the experts to define scenes as random occurrences (Sec. 5.3.2.7).

8.1.4 Causality Limitations

Pearl's Causal Theory as a mathematical theory is a relatively novel approach. The underling mathematical principles have been formalized in the last two decades. A lot of informal information about causality is available. The information also does not exist in a statistically digestible format. Its implementation to safety and especially to SOTIF domain is also novel. Pearl's Causal Theory as a formal language to model SOTIF provides new research horizon and also suffers from lack of formal notations for related statistical concepts.

Another important aspect presented in this thesis (and causality in general) is the use of causal and probability theory simultaneously. The commonly accepted definitions of causality revolve around a more deterministic realm while probability theory in general is used in the frequentist sense. In the proposed framework, the probability and causality are used together to define local and sometimes universal causal mechanisms. The limitation is that using these terms together has been critically reviewed by some researchers [23].

The causal framework presented oscillates between associational or purely statistical and interventional or purely causal mathematical language. The underlying assumption to justify such implementation is conversion of the causal quantities into associational ones for their mathematical calculation [87, 88]. This correspondence gives the liberty to use purely associational quantities while taking the causal assumption about the phenomena under study. However, it is the view of author that further robust axioms and postulates are required in order to enhance the confidence on the utilization of quantities with different mathematical semantics.

8.2 Methodological Level Limitations

8.2.1 Open Context Representation

Any modelling technique assumes the good approximation of the real world i.e., open context in this case. In the case of this thesis, the CBN represents the causal

model and the CBTs learnt through parameter learning represent the relative occurrence frequency as the approximation of the open world phenomena. It may happen that not all the causal factors are encoded in the CBN structure and dataset used are not generalizable. This results in error-prone maps (PLM, CPLM, and PPMI).

8.2.2 Resemblance to Structure Learning

A potential objection to the testable implication block's methodology can be its resemblance with the various structure learning methodologies. The block enhances the CBN structure by identifying novel triggering conditions. The methodology discussed in (Sec. 5.2.7.1) has the following distinct features.

1. In the testable implication block implementation, a human expert plays a major role in the construction of the methodology leading to an iterative safety analysis technique. It is reflected in the block (Fig. 5.1) and it is a missing feature of a structure learning technique.

2. The human expert in the loop analysis results in the identification of triggering conditions that may not be the part of available data. This indicates new CBN structure may acquire causal relations that are not present in the initial data. Structure learning is limited to whatever is available in the form of data.

8.3 Implementation Level Limitations

8.3.1 Results Generalization

In the previous chapter (Chap. 7), statements e.g., "Highly occluded scenes cause high FN" were deduced through the implementation of the causal framework provided in Chap. 5. In doing so, special attention was given to model the randomness in the causal relation. The implementation is based on dataset and its representative causal structure. However, in any data-oriented implementation, measuring the true parameter (or CBTs) is a challenging problem. A parameter learning algorithm of CBN calculates the joint relative frequency if they have conditional relation in their structure [66]. The method approximates the parameter (or CBTs) based on a dataset $\mathcal{D}$, as the real CBTs are unknown. These CBTs are representation of real-world distribution and are easily falsifiable by instance of unknown and distinct distributions. Tasks that are safety critical nature require extreme care in deducing such results.

This thesis addresses such shortcomings in two ways.

- It includes the human experts' inputs
- It dedicates a block for metrics to build confidence on the results

However, the generalization of results still requires community wide consensus.

8.3.2 Rare Event Occurrences

This limitation concerns the well documented problem of rare event occurrence in the safety analysis of a system and its context. It arises when some states of nodes occur with very low frequency e.g., *illumination=tunnel* and *truncation=yes* occur at a very low frequency in the dataset. These states can also be safety critical from the SOTIF standpoint. Evaluation of CPLMs for states with such dataset may lead to perturbation in the results. Such states can be artificially inserted in the data for better representation.

8.3.3 Train and Test Data

Train and test dataset affect both the plausibilization and testable implications block. Test dataset is inappropriately segregated from the training dataset. Generally, the two datasets should not be correlated. However, highly correlated ones are generally used as they are collected at the same location and time. This leads to overestimation in the accuracies of PLMs and CPLMs. Randomizing the selection of dataset (e.g., Monte Carlo) can assist in the optimization of the best solutions [58].

8.3.4 Availability of Labelled Data

Labelled data acquired during development is of paramount importance for the proposed framework. It is possible that data may not be available for some potential causal relation. For example, the potential causal relation with "car loaded on a trailer", "ground truth labelling error" and "construction activity" (Fig. 5.8) have no available data for the experiment performed. The unavailability of data and its acquisition can be addressed in the following ways.

1. *Labelling Automation:* Manual data labelling is error-prone, labour-intensive and expensive exercise. Part of labelling process can be automated by using labelling algorithms.

2. *Label Ranking:* In case of limited resources available for the labelling exercise, a ranking of labels based on some structure e.g., Phenomena Identification and Ranking Table (PIRT) [112] can be used.

9

Conclusion and Future Research

"It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment."
– Johann Carl Friedrich Gauss , *German Mathematician & Physicist*

In this chapter, the contribution of this thesis and a reflection on the research questions are presented (Sec. 9.1). This is followed by a set of suggestions for future research (Sec. 9.2).

9.1 Conclusion

In Ch. 1, three research questions are introduced.

Research Question 1: Can existing safety analysis techniques such as FTA/ FMEA model SOTIF/ safety of the automated driving? If not, what aspects can they not model?

This question is discussed in Ch. 3. First, challenging safety aspects of automated driving are discussed by providing a system theoretic view of the problem. It is then followed by a comprehensive critical view on the safety analyses techniques referenced by ISO 21448 [54] to provide SOTIF analysis. The SOTIF analysis of HAD vehicles' situational awareness is considered in this dissertation.

HAD vehicles are complex systems operating in open context. The complexity and open context nature may result in unsafe and uncertain behaviour due to triggering conditions and insufficiencies in sensing and understanding the operational environment. Specifically, the inherent randomness is present in the occurrences and causal relations of phenomena. There may also exist complex causal interactions between different phenomena leading to a spurious relation between dependent and independent variables. Moreover, a lack of knowledge may also be present regarding the existence of triggering conditions that effects performance limitations, their occurrences and causal relations (Sec. 3.3). The Ch. 3 also associates randomness, variability and lack of knowledge to uncertainty.

ISO 21448 references FTA, FMEA and STPA to perform safety analyses for automated driving functions. The methods are highly effective in the automotive industry to assess functional safety; however, they do not model randomness and lack of knowledge about triggering conditions. Moreover, complex causal phenomena cannot be modelled using these techniques as well (Sec. 3.4.3). The concepts of causations are based on the assumptions that are seldom fulfilled. Another aspect of safety that emanates from HAD vehicle operating context is the knowledge of all

the triggering conditions on which the performance of the HAD vehicle is conditioned. Novel triggering conditions may be discovered during testing phases if data is collected and analysed. Traditional safety analysis techniques lack the framework to discover, identify and incorporate novel triggering conditions in the existing models.

Research Question 2: Which safety analysis models are suitable to represent different type/facets of uncertainties encountered in complex systems and open context?

This question is discussed in Ch. 4. First a taxonomy of uncertainty is provided followed by formulation of the safety analyses models that can represent uncertainty.

This chapter introduces three major categorizations of uncertainty i.e., aleatory, epistemic and ontological uncertainty. These uncertainties represent randomness, lack of knowledge and complete state of ignorance, respectively.

In literature, uncertainty is generally categorized as aleatory or epistemic. However, a further distinction into ontological uncertainty provides a valuable aspect to represent uncertainties in safety analysis models. Ontological uncertainty requires a different means for treatment. This amounts to modelling the novel triggering conditions into the safety analysis as they are discovered through iterative process. For complex systems operating in an open context the ontological uncertainty can never be completely disregarded. Therefore, it should be included in the safety case so that it has been properly addressed and represented.

The chapter also introduces BN, CBN, EN and EEN with the ability to model one or all the categories of uncertainties, which provides an answer to research question 2. Although the networks introduced provide a comprehensive representation of the subjective interpretation of uncertainties, epistemic uncertainty lacks in providing the semantics when measurement metric is involved.

In the concluding section of the Ch. 4, a SOTIF analysis using CBN and EEN is demonstrated. The demonstration example shows how SOTIF relevant triggering conditions affect the FN probability. In the case of CBN, only one measure is used that represents aleatory uncertainty. Relevant SOTIF improvement measure can be dictated based on this measure e.g., better perception system with lower FN distribution values. The EEN provides four different measures which can be used to represent aleatory, epistemic and ontological uncertainty. One extra measure represents the coupling between epistemic and ontological uncertainty. The expert can provide subjective assessment of ontological uncertainty based on the current knowledge about the system and open context.

The third research question explores the safety analysis models that can support an augmentation of the safety analysis iteratively.

Research Question 3: How can safety analysis models be applied to support an iterative augmentation of the safety analysis and enable discovery of new knowledge encountered in complex systems and open context?

This research question is addressed in Ch. 5 of this dissertation while supporting case study implementation and evaluation is provided in Ch. 6 and Ch. 7, respectively. Ch. 5 provides details of a high-level framework followed by two specific algorithms about the implementation.

The framework introduced is a generic abstraction with the ability to inculcate various mathematical implementations. The framework underpins two important

aspects while also providing uncertainty representation: (1) A causal model that provides a representation to triggering conditions and their effects on performance limitation. (2) Ability to systemically discover, identify and model novel triggering conditions. These aspects are modelled by CBN structure, conditional probabilities as well as p-value hypothesis testing applied on the CBN and test datasets discussed in detail in the representative algorithms.

The framework together with its algorithms not only addresses the shortcomings of the safety analyses techniques detailed in Ch. 3, it also provides explanation of performance limitations, extraction of relevant triggering conditions, derivation of an abstract open context model, scenario catalogization and causal effect evaluation. It supports a systematic refinement loop to enhance the confidence in the modelled causal relations. The framework also supports iterations to assist in the semi-automated discovery of triggering conditions. The identified triggering conditions can be used to define SOTIF improvement measures while the scenario catalogization and ODD model derived from the implementation of the framework can define direction for V&V tests.

In order to support this research question, a case study based on real world dataset about the LIDAR detections is evaluated. The evaluation supports the results through an initial causal model in which multiple relevant triggering conditions are identified along with the augmentation of the model through a novel triggering condition discovered in iteration.

9.1.1 Summary

To conclude, safety analysis techniques established in the automotive industry are analysed in this thesis. Novel safety analysis methods that incorporate multifaceted uncertainties in their models are introduced to model SOTIF for HAD functions. These techniques enhance the modelling capability of safety analyses and provide more meaningful semantics to safety models. More importantly, the introduced novel techniques assist in SOTIF analysis of HAD vehicles operating in open context. The thesis also provides a comprehensive framework that not only models the uncertainties, it also provides an iterative approach to augment the SOTIF analysis models by introducing systematic discovery of novel triggering conditions. The framework supports a hybrid approach i.e., an expert and data engineering-oriented approach. The proposed framework is validated through a real-world case study in which dataset of LIDAR based cars' detection is used. Limitation of the approach, robustness concerns and threat to validity are also discussed in detail.

9.2 Future Research

The probabilistic modelling approaches for SOTIF is in its nascent research stage. With the aim that this dissertation provides a way forward to future research, important future research topics are discussed in the following.

First, theoretical aspects of the proposed approach that can be the theme of future research are discussed. This covers the limitations and threat to validity measures discussed in Ch. 8. Afterwards, an insight on the possible research directions on the practical implementations is provided, followed by the propositions on the extension of the results.

Categorization of uncertainties into aleatory, epistemic and ontology for SOTIF is a novel approach. In this dissertation ad-hoc modelling of CBN is used to represent EN and EEN (as an extension to CBN). In the literature, graphs supporting Evidence

Theory have been proposed [113, 107]. Graphical techniques based on these publications should also be proposed for EEN.

In this dissertation, the plausibilization of a causal effect is either defined based on the expert opinion or hypothetical threshold values. This study could be extended by the incorporation of the formalised plausibilization based on the principles defined in the regulation and management of safety-critical and safety-involved systems e.g., As Low As Reasonably Possible (ALARP) or Globalement Au Moins Aussi Bon (GAMAB). Influence diagrams can be considered as a natural representation of causal network that include decision-making [57]. Influence diagrams provide an added benefit in placing a formalised value on information.

Increasing standardization of the processes involved in defining the abstractions e.g., through structured descriptions of scenarios [101] at which the CBN is constructed will benefit the industry-wide acceptance of the CBN and the framework. Formalization of the traffic and environment by using a structured description such as defined by Scholtes [101] can assist in defining consistent CBN structures.

In this study, a single PLM was used to define the performance of the perception system. In real world applications, multiple PLMs need to be used to argue the adequacy of SOTIF. A formal methodology can be devised to select a manageable set of PLMs.

In this dissertation, a specific case was chosen for mathematical implementation. However, the proposed framework is generic in nature and multiple mathematical implementations can be provided. A thorough comparative analysis is required showcasing the implementations and variation in results. For example, the p-value hypothesis testing can be replaced by other inferential statistic techniques [108], to further analyse the mathematical aspects of the proposed framework as future research.

Safety dashboards are slowly becoming an integral part of the safety engineering process. They provide a seamless integration of expert knowledge and data analytics and may result in scenario catalogue and assist the expert in identification of triggering conditions. The proposed framework in this dissertation can be used to implement such dashboards. It may also speed up the process of data processing and analytics for safety engineering practices.

Bibliography

[1] Ahmad Adee, Roman Gansch, and Peter Liggesmeyer. "Systematic Modeling Approach for Environmental Perception Limitations in Automated Driving". In: *2021 17th European Dependable Computing Conference (EDCC)*. 2021, pp. 103–110. DOI: 10.1109/EDCC53658.2021.00022.

[2] Ahmad Adee et al. "Discovery of Perception Performance Limiting Triggering Conditions in Automated Driving". In: *2021 5th International Conference on System Reliability and Safety (ICSRS)*. 2021, pp. 248–257. DOI: 10.1109/ICSRS53853.2021.9660641.

[3] Ahmad Adee et al. "Uncertainty Representation with Extended Evidential Networks for Modeling Safety of the Intended Functionality (SOTIF)". In: *European Safety and Reliability Conference (ESREL2020)*. 2020, pp. 4148–4156.

[4] Harish Agarwal et al. "Uncertainty quantification using evidence theory in multidisciplinary design optimization". In: *Reliability Engineering & System Safety* 85.1-3 (2004), pp. 281–294.

[5] Felipe Aguirre et al. "Application of evidential networks in quantitative analysis of railway accidents". In: *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* 227.4 (2013), pp. 368–384.

[6] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. "A survey of network anomaly detection techniques". In: *Journal of Network and Computer Applications* 60 (2016), pp. 19–31.

[7] N. Ali, M. Hussain, and J. E. Hong. "Analyzing Safety of Collaborative Cyber-Physical Systems Considering Variability". In: *IEEE Access* 8 (2020), pp. 162701–162713. DOI: 10.1109/ACCESS.2020.3021460.

[8] Qazi Muhammad Nouman Amjad, Muhammad Zubair, and Gyunyoung Heo. "Modeling of common cause failures (CCFs) by using beta factor parametric model". In: *2014 International Conference on Energy Systems and Policies (ICESP)*. IEEE. 2014, pp. 1–6.

[9] Algirdas Avizienis et al. "Basic Concepts and Taxonomy of Dependable and Secure Computing". In: *IEEE Trans. Dependable Sec. Comput.* 1.1 (2004), pp. 11–33.

[10] Jie Bai et al. "Application Oriented Identification of External Influence Factors on Sensing for Validation of Automated Driving Systems". In: *2019 IEEE MTT-S International Wireless Symposium (IWS)*. 2019, pp. 1–3. DOI: 10.1109/IEEE-IWS.2019.8803852.

[11] Boutheina Bannour, Julien Niol, and Paolo Crisafulli. "Symbolic Model-based Design and Generation of Logical Scenarios for Autonomous Vehicles Validation". In: *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2021, pp. 215–222.

[12] Leonard Baum, Tom Assmann, and Henning Strubelt. "State of the art-Automated micro-vehicles for urban logistics". In: *IFAC-PapersOnLine* 52.13 (2019), pp. 2455–2462.

[13] Christopher Becker, John C Brewer, Larry Yount, et al. *Safety of the intended functionality of lane-centering and lane-changing maneuvers of a generic level 3 highway chauffeur system*. Tech. rep. United States. National Highway Traffic Safety Administration. Electronic . . ., 2020.

[14] Mario Berk. "Safety Assessment of Environment Perception in Automated Driving Vehicles". PhD thesis. Technische Universität München, 2019.

[15] Laura E Bothwell and Scott H Podolsky. "The emergence of the randomized, controlled trial". In: *N Engl J Med* 375.6 (2016), pp. 501–504.

[16] Marco Bozzano and Adolfo Villafiorita. *Design and safety assessment of critical systems*. CRC press, 2010.

[17] British Standard Institute. "BSI/ PAS 1883: 2020 Operational Design Domain (ODD) taxonomy for an automated driving system (ADS) - Specification". In: (2020), p. 285.

[18] Simon Burton. *A causal model of safety assurance for machine learning*. 2022. DOI: 10.48550/ARXIV.2201.05451. URL: https://arxiv.org/abs/2201.05451.

[19] Simon Burton et al. "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective". In: *Artificial Intelligence* 279 (2020), p. 103201.

[20] Simon Burton et al. "Safety, Complexity, and Automated Driving: Holistic Perspectives on Safety Assurance". In: *Computer* 54.8 (2021), pp. 22–32.

[21] Baoping Cai et al. "Application of Bayesian networks in reliability evaluation". In: *IEEE Transactions on Industrial Informatics* 15.4 (2018), pp. 2146–2157.

[22] Brigitte Chaput, Jean-Claude Girard, and Michel Henry. "Frequentist approach: Modelling and simulation in statistics and probability teaching". In: *Teaching Statistics in school mathematics-Challenges for teaching and teacher education*. Springer, 2011, pp. 85–95.

[23] Lauchlan J Clarke. "A resilience-based causal framework for conducting safety analysis". PhD thesis. University of Tasmania, 2020.

[24] John Clarkson and Claudia Eckert. *Design process improvement: a review of current practice*. Springer Science & Business Media, 2010.

[25] Ethics Commission et al. "Automated and connected driving". In: *Federal ministry of transport and digital infrastructure, Germany. Available online at: https://www.bmvi. de/SharedDocs/EN/publications/report-ethics-commission. html* (2017).

[26] Paul Cooper. "Data, information, knowledge and wisdom". In: *Anaesthesia & Intensive Care Medicine* 18.1 (2017), pp. 55–56.

[27] Erwin De Gelder et al. "Risk Quantification for Automated Driving Systems in Real-World Driving Scenarios". In: *IEEE Access* 9 (2021), pp. 168953–168970.

[28] Arthur P Dempster. "A generalization of Bayesian inference". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 30.2 (1968), pp. 205–232.

[29] Yao Deng et al. "An analysis of adversarial attacks and defenses on autonomous driving models". In: *2020 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE. 2020, pp. 1–10.

[30] Armen Der Kiureghian and Ove Ditlevsen. "Aleatory or epistemic? Does it matter?" In: *Structural Safety* 31.2 (2009), pp. 105–112.

[31] Klaus Dietmayer. "Predicting of machine perception for automated driving". In: *Autonomous Driving*. Springer, 2016, pp. 407–424.

[32] Suraj R Dwivedi et al. "Intelligent Traffic Management System". In: (2020).

[33] John Earman et al. *A primer on determinism*. Vol. 37. Springer Science & Business Media, 1986.

[34] Péter Érdi. *Complexity explained*. Springer, 2008.

[35] Norman Fenton and Martin Neil. *Risk assessment and decision analysis with Bayesian networks*. Crc Press, 2018.

[36] Scott Ferson et al. *Constructing probability boxes and Dempster-Shafer structures*. Tech. rep. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2015.

[37] M Julia Flores et al. "Incorporating expert knowledge when learning Bayesian network structure: a medical case study". In: *Artificial intelligence in medicine* 53.3 (2011), pp. 181–204.

[38] Linton C Freeman. "Elementary Applied Statistics. New York: JohnWiley and Sons". In: *Inc., 19t* 6 (1965).

[39] Chao Fu, Jian-Bo Yang, and Shan-Lin Yang. "A group evidential reasoning approach based on expert reliability". In: *European Journal of Operational Research* 246.3 (2015), pp. 886–893.

[40] Roman Gansch. "System theoretic approach for uncertainty in safety analysis of automated driving". PhD thesis. June 2019.

[41] Roman Gansch and Ahmad Adee. "System Theoretic View on Uncertainties". In: *2020 Design, Automation Test in Europe Conference Exhibition (DATE)*. 2020, pp. 1345–1350. DOI: 10.23919/DATE48585.2020.9116472.

[42] Erwin de Gelder, A Khabbaz Saberi, and Hala Elrofai. "A method for scenario risk quantification for automated driving systems". In: *26th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*. Mira Smart. 2019.

[43] Alan E Gelfand. "Gibbs sampling". In: *Journal of the American statistical Association* 95.452 (2000), pp. 1300–1304.

[44] Hadi Ghahremannezhad, Hang Shi, and Chenajun Liu. "Robust road region extraction in video under various illumination and weather conditions". In: *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*. IEEE. 2020, pp. 186–191.

[45] Christopher Goodin et al. "Predicting the Influence of Rain on LIDAR in ADAS". In: *Electronics* 8.1 (2019), p. 89.

[46] Sorin Grigorescu et al. "A survey of deep learning techniques for autonomous driving". In: *Journal of Field Robotics* 37.3 (2020), pp. 362–386.

[47] Teemu Hakala et al. "Spectral imaging from UAVs under varying illumination conditions". In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. International Society for Photogrammetry and Remote Sensing (ISPRS). 2013.

[48] Lars Harms-Ringdahl. *Guide to safety analysis for accident prevention*. Citeseer, 2013.

[49] Daniel Hastings and Hugh McManus. "A framework for understanding uncertainty and its mitigation and exploitation in complex systems". In: *Engineering Systems Symposium*. 2004, pp. 29–31.

[50] Marcus Hedlund. *Weather Influence on LiDAR Signals using the Transient Radiative Transfer and LiDAR Equations*. 2020.

[51] Francis Heylighen and Cliff Joslyn. "Cybernetics and second-order cybernetics". In: *Encyclopedia of physical science & technology* 4 (2001), pp. 155–170.

[52] Altunkan Hizal. "Improving performance of an FMCW radar in rain and sea clutter by satistical dithering". In: *IET International Radar Conference 2015*. IET. 2015, pp. 1–5.

[53] Manzoor Hussain et al. "Analyzing Safety in Collaborative Cyber-Physical Systems: A Platooning Case Study". In: ().

[54] International Organization for Standardization. "ISO 21448:2021 Road vehicles – Safety of the intended functionality". In: (2021), p. 181.

[55] International Organization for Standardization. *ISO 26262-1:2011 Road vehicles — Functional safety*. ISO 26262. Norm. ISO, Geneva, Switzerland, 2011.

[56] International Organization for Standardization. "ISO 4804:2020 Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation". In: (2020), p. 125.

[57] Frank Jensen, Finn V. Jensen, and Søren L. Dittmer. "From Influence Diagrams to Junction Trees". In: *Uncertainty Proceedings 1994*. Ed. by Ramon Lopez de Mantaras and David Poole. San Francisco (CA): Morgan Kaufmann, 1994, pp. 367–373. ISBN: 978-1-55860-332-5. DOI: https://doi.org/10.1016/B978-1-55860-332-5.50051-1. URL: https://www.sciencedirect.com/science/article/pii/B9781558603325500511.

[58] Hao Ji and Yaohang Li. "Monte Carlo methods and their applications in Big Data analysis". In: *Mathematical Problems in Data Science*. Springer, 2015, pp. 125–139.

[59] Zhiwei Ji, Qibiao Xia, and Guanmin Meng. "A review of parameter learning methods in Bayesian network". In: *International Conference on Intelligent Computing*. Springer. 2015, pp. 3–12.

[60] Ratiba Kabli, Frank Herrmann, and John McCall. "A chain-model genetic algorithm for Bayesian network structure learning". In: *Proceedings of the 9th annual conference on genetic and evolutionary computation*. 2007, pp. 1264–1271.

[61] Janusz Kacprzyk. *New Metaheuristics, Neural and Fuzzy Techniques in Reliability [electronic Resource]*. Springer.

[62] Nidhi Kalra and Susan M Paddock. "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" In: *Transportation Research Part A: Policy and Practice* 94 (2016), pp. 182–193.

[63] Marzana Khatun, Michael Glaß, and Rolf Jung. "A Systematic Approach of Reduced Scenario-based Safety Analysis for Highly Automated Driving Function." In: *VEHITS*. 2021, pp. 301–308.

[64] Marzana Khatun, Michael Glaß, and Rolf Jung. "Scenario-based extended hara incorporating functional safety and sotif for autonomous driving". In: *Proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference*. 2020, pp. 53–59.

[65] OM Kirovskii and VA Gorelov. "Driver assistance systems: analysis, tests and the safety case. ISO 26262 and ISO PAS 21448". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 534. 1. IOP Publishing. 2019, p. 012019.

[66] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[67] Birte Kramer et al. "Identification and Quantification of Hazardous Scenarios for Automated Driving". In: *Model-Based Safety and Assessment*. Ed. by Marc Zeller and Kai Höfig. Cham: Springer International Publishing, 2020, pp. 163–178.

[68] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79 –86. DOI: 10.1214/aoms/1177729694. URL: https://doi.org/10.1214/aoms/1177729694.

[69] Pierre-Simon Laplace. *Pierre-Simon Laplace philosophical essay on probabilities: translated from the fifth french edition of 1825 with notes by the translator*. Vol. 13. Springer Science & Business Media, 1998.

[70] Nancy Leveson and JOHN Thomas. "STPA handbook". In: *Partnership for Systems Approaches to Safety and Security (PSASS)* 3 (2018).

[71] Nancy Leveson et al. "Engineering resilience into safety-critical systems". In: *Resilience engineering*. CRC Press, 2017, pp. 95–123.

[72] Michel Loeve. *Probability theory*. Courier Dover Publications, 2017.

[73] Yasser A Mahmood et al. "Fuzzy fault tree analysis: a review of concept and application". In: *International Journal of System Assurance Engineering and Management* 4.1 (2013), pp. 19–32.

[74] Helmut Martin et al. "Identification of performance limitations of sensing technologies for automated driving". In: *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE)*. IEEE. 2019, pp. 1–6.

[75] Edward McFowland, Skyler Speakman, and Daniel B Neill. "Fast generalized subset scan for anomalous pattern detection". In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 1533–1561.

[76] Edward McFowland III et al. "Automated Discovery of Novel Anomalous Patterns". In: ().

[77] Raymond J Mikulak, Robin McDermott, and Michael Beauregard. *The basics of FMEA*. CRC press, 2017.

[78] Robert P Murphy. *Chaos theory*. Ludwig von Mises Institute, 2010.

[79] Shree K Nayar and Srinivasa G Narasimhan. "Vision in bad weather". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. IEEE. 1999, pp. 820–827.

[80] Christian Neurohr et al. "Criticality Analysis for the Verification and Validation of Automated Vehicles". In: *IEEE Access* 9 (2021), pp. 18016–18041.

[81] Richard Ni and Jason Leung. "Safety and Liability of Autonomous Vehicle Technologies". In: *Massachusetts Institute* (2014).

[82] Xiaojuan Ning et al. "An efficient outlier removal method for scattered point cloud data". In: *PloS one* 13.8 (2018), e0201280.

[83] HC Oliveira. "Occlusion detection by height gradient for true orthophoto generation, using LiDAR data". In: (2013).

[84] Bruno A Olshausen. "Bayesian probability theory". In: *The Redwood Center for Theoretical Neuroscience, Helen Wills Neuroscience Institute at the University of California at Berkeley, Berkeley, CA* (2004).

[85] Christina Pakusch et al. "Unintended effects of autonomous driving: A study on mobility preferences in the future". In: *Sustainability* 10.7 (2018), p. 2404.

[86] Riccardo Patriarca et al. "The past and present of System-Theoretic Accident Model And Processes (STAMP) and its associated techniques: A scoping review". In: *Safety Science* 146 (Feb. 2022), p. 105566. DOI: 10.1016/j.ssci.2021.105566.

[87] Judea Pearl. *Causality*. Cambridge university press, 2009.

[88] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[89] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[90] Sergio Peñafiel et al. "Associating risks of getting strokes with data from health checkup records using Dempster-Shafer Theory". In: *2018 20th International Conference on Advanced Communication Technology (ICACT)*. 2018, pp. 239–246. DOI: 10.23919/ICACT.2018.8323710.

[91] Alexander Poddey et al. "On the validation of complex systems operating in open contexts". In: *arXiv preprint arXiv:1902.10517* (2019).

[92] Uwe Kay Rakowsky. "Fundamentals of the Dempster-Shafer theory and its applications to reliability modeling". In: *international journal of Reliability, Quality and Safety Engineering* 14.06 (2007), pp. 579–601.

[93] Raymond Reiter. "On closed world data bases". In: *Readings in artificial intelligence*. Elsevier, 1981, pp. 119–140.

[94] Stefan Riedmaier et al. "Survey on scenario-based safety assessment of automated vehicles". In: *IEEE Access* 8 (2020), pp. 87456–87477.

[95] Francois Role and Mohamed Nadif. "Handling the impact of low frequency events on co-occurrence based measures of word similarity". In: *Proceedings of the international conference on Knowledge Discovery and Information Retrieval (KDIR-2011). Scitepress*. 2011, pp. 218–223.

[96] Robert Rosen. *Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life*. Columbia University Press, 1991.

[97] SAE International. *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. 2014.

[98] Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. "A survey on Bayesian network structure learning from data". In: *Progress in Artificial Intelligence* 8.4 (2019), pp. 425–439.

[99] Adam Schnellbach and Gerhard Griessnig. "Development of the ISO 21448". In: *European Conference on Software Process Improvement*. Springer. 2019, pp. 585–593.

[100] Maike Scholtes and Lutz Eckstein. *Systematic Categorization of Influencing Factors on Radar-Based Perception to Facilitate Complex Real-World Data Evaluation*. 2021. DOI: 10.48550/ARXIV.2105.00279. URL: https://arxiv.org/abs/2105.00279.

[101] Maike Scholtes et al. "6-layer model for a structured description and categorization of urban traffic and environment". In: *IEEE Access* 9 (2021), pp. 59131–59147.

[102] Lisa Schut et al. "Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1756–1764.

[103] E. Schwalb. "Analysis of Safety of The Intended Use (SOTIF)". In: 2019.

[104] Glenn Shafer. *A mathematical theory of evidence*. Vol. 42. Princeton university press, 1976.

[105] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. "On a formal model of safe and scalable self-driving cars". In: *arXiv preprint arXiv:1708.06374* (2017).

[106] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[107] Prakash P. Shenoy. "Valuation-based systems: a framework for managing uncertainty in expert systems". In: 1992.

[108] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC, 2003.

[109] Christophe Simon and Philippe Weber. "Evidential networks for reliability analysis and performance evaluation of systems with imprecise knowledge". In: *IEEE Transactions on Reliability* 58.1 (2009), pp. 69–87.

[110] Christophe Simon, Philippe Weber, and Alexandre Evsukoff. "Bayesian networks inference algorithm to implement Dempster Shafer theory in reliability analysis". In: *Reliability Engineering & System Safety* 93.7 (2008), pp. 950–963.

[111] Christophe Simon, Philippe Weber, and Eric Levrat. "Bayesian networks and evidence theory to model complex systems reliability". In: (2007).

[112] Preet M Singh et al. "Phenomena Identification and Ranking Table (PIRT) study for metallic structural materials for advanced High-Temperature reactor". In: *Annals of Nuclear Energy* 123 (2019), pp. 222–229.

[113] Philippe Smets. "Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem". In: *Int. J. Approx. Reasoning* 9.1 (1993), pp. 1–35.

[114] David J Smith and Kenneth GL Simpson. *Safety critical systems handbook: a straight forward guide to functional safety, IEC 61508 (2010 Edition) and related standards, including process IEC 61511 and machinery IEC 62061 and ISO 13849*. Elsevier, 2010.

[115] Ian Sommerville. *Engineering software products*. Pearson London, 2020.

[116] Bernd Spanfelner et al. "Challenges in applying the ISO 26262 for driver assistance systems". In: *Tagung Fahrerassistenz, München* 15.16 (2012).

[117] Diomidis H Stamatis. *Failure mode and effect analysis: FMEA from theory to execution*. ASQ Quality press, 2003.

[118] Susan Stepney. "Complex systems for narrative theorists". In: *Narrating complexity*. Springer, 2018, pp. 27–36.

[119] Alicia Sudol. "A methodology for modeling the verification, validation, and testing process for launch vehicles". PhD thesis. Georgia Institute of Technology, 2015.

[120] Robert Sugden. "Spontaneous order". In: *Journal of Economic perspectives* 3.4 (1989), pp. 85–97.

[121] Nassim Nicholas Taleb and Mark Blyth. "The black swan of Cairo: How suppressing volatility makes the world less predictable and more dangerous". In: *Foreign Affairs* (2011), pp. 33–39.

[122] Stephen Thomas and Katrina M Groth. "Toward a hybrid causal framework for autonomous vehicle safety analysis". In: *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability* (2021), p. 1748006X211043310.

[123] Underwriters Laboratories. "UL 4600: 2020 Safety For The Evaluation Of Autonomous Products". In: (2020), p. 285.

[124] Lev V Utkin and Frank PA Coolen. "Imprecise reliability: an introductory overview". In: *Computational intelligence in reliability engineering* (2007), pp. 261–306.

[125] Faruk Uysal and Sasanka Sanka. "Mitigation of automotive radar interference". In: *2018 IEEE Radar Conference (RadarConf18)*. IEEE. 2018, pp. 0405–0410.

[126] John P Van Gigch. *System design modeling and metamodeling*. Springer Science & Business Media, 1991.

[127] Sebastian Vander Maelen et al. "An Approach for Safety Assessment of Highly Automated Systems Applied to a Maritime Traffic Alert and Collision Avoidance System". In: *2019 4th International Conference on System Reliability and Safety (ICSRS)*. IEEE. 2019, pp. 494–503.

[128] M Alex O Vasilescu, Eric Kim, and Xiao S Zeng. "CausalX: Causal eXplanations and block multilinear factor analysis". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 10736–10743.

[129] William E Vesely et al. *Fault tree handbook*. Tech. rep. Nuclear Regulatory Commission Washington DC, 1981.

[130] Nguyen Xuan Vinh et al. "Data discretization for dynamic Bayesian network based modeling of genetic networks". In: *International Conference on Neural Information Processing*. Springer. 2012, pp. 298–306.

[131] Peter Walley. "Statistical reasoning with imprecise probabilities". In: (1991).

[132] Philippe Weber et al. "Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas". In: *Engineering Applications of Artificial Intelligence* 25.4 (2012), pp. 671–682.

[133] Wilhard Wendorff. "Quantitative SOTIF Analysis for highly automated Driving Systems". In: Nov. 2017.

[134] Zhitao Wu et al. "Multi-Scale Software Network Model for Software Safety of the Intended Functionality". In: *2021 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE. 2021, pp. 250–255.

[135] Shijie Zhang, Tao Tang, and Jintao Liu. "A hazard analysis approach for the sotif in intelligent railway driving assistance systems using stpa and complex network". In: *Applied Sciences* 11.16 (2021), p. 7714.

[136] Xinhai Zhang et al. “Finding critical scenarios for automated driving systems: A systematic literature review”. In: *arXiv preprint arXiv:2110.08664* (2021).

[137] Naaman Zhou. “Volvo admits its self-driving cars are confused by kangaroos”. In: *The Guardian* (2017).

Ahmad **Adee**

SOTIF Expert

linkedin.com/in/adee4102 Munich, Germany

Professional Experience

Present – July 2023

SOTIF Expert | BMW AG, Germany

As a SOTIF expert the main tasks revolve around SOTIF analysis, verification and validation of automated driving functions.

Python | ISO 21448 | ISO PAS 8800

June 2023 – December 2021

Senior Safety Engineer | TomTom N.V., The Netherlands

As a safety engineer the main tasks revolve around functional safety, SOTIF and AI safety activities.

- Provide HARA, top down and safety analyses for products including ADAS products
- Lead the execution of the safety framework for reliable map attributes as per ISO 5083
- Produce insufficiencies of specifications for reliable map attributes to address SOTIF requirements
- Communicate with the product teams for the safety by design development of the products in line with ISO 26262
- Initiative on developing error propagation model (statistical analysis) of map production process to perform risk evaluation
- Leading the development of AI safety framework for company's product in the light of ISO PAS 8800, ISO 5469 and other relevant AI safety standards
- Developing framework for data intelligence implementation for product safety
- Provision of safety strategies as response to safety requirements requested in OEMs RFQs
- Defined tool qualification strategy for map generation process as per ISO 26262

Python | SOX | Netica | ISO 26262 | ISO 21448 | UL 4600 | ISO 5083 | ISO PAS 8800

November 2021 – February 2019

EU Researcher | Robert Bosch GmBH, Germany

The research revolves around developing safety analysis methods to assure safety of autonomous systems deployed in the complex environment. Development of these method are in line with ISO 21448. In return, these methods provide valuable insights on the safety architecture and safety requirements while also providing valuable inputs for the safety case generation. The developed methods are then implemented on practical applications. The salient features of the research are.

- Knowledge in probability theory, statistics, and quantitative risk analysis
- Definition of top level SOTIF improvement measures and requirements (SOTIF safety concept)
- Safety analysis as well as V&V method for safety of the intended functionality based on statistical methods and data analytics
- Identification and discovery of SOTIF related hazards, triggering conditions and performance limitations so as to limit unknown unsafe scenarios
- Argumentation on safety and uncertainties modeling using Bayesian network (BN)
- Representation of uncertainty and how uncertainty can represent SOTIF using probabilistic graphical models (PGMs) and BN
- Safety data analysis through data mining/ statistical methods used for BNs to identify hazards, triggering conditions and ODD related external disturbances (weather and road conditions etc).
- Argumentation on lack of knowledge and state of ignorance in PGMs and BNs
- Providing equivalency between FTA and BN
- Novel triggering condition/ hazard identification at scene/ scenario level by statistical hypothesis testing
- Supporting internal stakeholders and business units for risk assessment/ technology transfer

Python | Matlab | Simulink | Pandas | ROS | Carla | Bayes Server | Netica

January 2020 – November 2019

Visiting Researcher | Tecnische Universitat Kaiserslautern, Germany

As the visiting researcher for three months in the university, the following aspect was the focus of the research.

- Development of safety method for implementation of the SOTIF.

Python | Bayes Server | Netica

August 2018 – February 2018 | **Master Thesis Student | The laboratory of Digital Sciences of Nantes (LS2n), France**
Accuracy Improvement of Under-actuated 1-DOF (6-joints) mechanism : The mechanism is designed for calibration of industrial robots with a more efficient methodology than current industrial practice.
- Geometric Calibration of Under-actuated 1-DOF (6-joints) mechanism
- Stiffness Analysis of Under-actuated 1-DOF (6-joints) mechanism

Matlab | Simulink | Kuka simulator | Leica

September 2016 – September 2015 | **Assistant Production Engineer | MOL Pakistan Oil and Gas Co. B.V., Pakistan**
- Understanding of systems DCS Centum CS3000 Yokogawa, Honeywell, ESD HIMA
- Installation, calibration and loop verification of the field instruments
- Familiar with engineering standards ISA, IEEE, ASTM, NAMUR, ASMI, ANSI, NEMA, IEC

IBM Maximo | KELTON Flocal | Flow Computer floboss s600 | RSLogix 5000 | MON2000 Gas Chromatograph

August 2015 – September 2014 | **Trainee Engineer | MOL Pakistan Oil and Gas Co. B.V., Pakistan**
As a Trainee Engineer, I was a part of Growww 2014 training program conducted by MOL Group, Hungary. During this tenure, not only I got hands on experience in the field but also gained experience under professional guidance and developed an instinct to work in cross cultural working environment.
IBM Maximo | KELTON Flocal | Flow Computer floboss s600 | AMATEK 3050 Customer Configuration | RSLogix 5000

September 2013 – June 2014 | **Bachelor Thesis Student | National University of Sciences and Technology, Pakistan**
Design, Analysis, Fabrication & Control of an Unmanned Ground Vehicle that is throw able, invertable and can be used for reconnaissance and surveillance.
Arduino | Matlab | Solidworks

Education

September 2017 - August 2018	European **Masters** in Advance Robotics | Ecole Centrale de Nantes, France
September 2016 - July 2017	European **Masters** in Advance Robotics | Warsaw University of Technology, Poland
September 2010 - June 2014	**Bachelor** of Mechatronics Engineering| National University of Sciences and Technology, Pakistan

Standard Organizations

2022-2023	ISO/IEC JTC 1/SC 42/WG 2 ”Data”, Member
2022-2023	ISO/IEC JTC 1/SC 42/WG 3 ”Trustworthiness”, Member
2022-2023	ISO TC22/SC32/WG 14 ”Safety and Artificial Intelligence”, Member

Interpersonal/Organizational Skills & Honor/Awards

Planning | Priortization | Time Management | Leadership | Marie Curie Fellowship | Erasmus Mundus Scholarship