Testing Automated Driving Systems To Calibrate Drivers' Trust

By

Siddartha Khastgir

A thesis submitted in partial fulfilment of the requirements of the degree of

Doctor of Philosophy in Engineering

University of Warwick, Warwick Manufacturing Group (WMG)

May 2019

Dedicated to my Dadu.

Thank you for laying the foundation for everything I know today.

I owe this to you.

TABLE OF CONTENTS

ACKNOWLEDGEMENT

My PhD journey would not have been as enjoyable and rewarding without the unwavering support and guidance from many people. Firstly, I want to thank Professor Paul Jennings for believing in me, inspiring me and at the same time continuously challenging me to keep improving. For your selfless mentoring and guidance, I will forever be indebted to you. I am ever so grateful for your mentorship right from my undergraduate days since that fateful summer internship in 2009.

I would like to thank Dr Stewart Birrell for the constant encouragement and help with academic writing and publications. Your detailed inputs, hugely improved the driving simulator study designs. Thank you for being there to discuss various aspects of the research and guiding me to bring the two themes of this research together. Additionally, I am hugely grateful to Mr. Gunwant (Gunny) Dhadyalla for always being there to discuss my research ideas and their impact. I was very fortunate to have three mentors who have also been on this journey with me.

This research would not have been possible without my Mr. Administrative Magician – Jonathan Smith. I am grateful for your support and ability to resolve all administrative deadlocks. This research would not have had the desired impact without your backing and the Guinnesses we shared!

I also want to extend my sincere gratitude to Simon Brewerton from RDM Group. Without your support and inputs, the work on STPA of Low-Speed Automated Driving System would not have been realised.

Dr Erik Kempert, with whom I have shared many laughs and had indepth discussions about everything on the planet, made the long hours in the simulator enjoyable. Thank you Erik for everything. I hugely admire you. Any simulator study at WMG would not be possible without the support from Dr Jakes Groenewald. Thank you Jakes for ensuring that I and all other research students always had an excellent simulator support to work with.

I would also like to thank my other colleagues and researchers with whom I have spent countless hours in the simulator and in discussions. Thank you Vadim, Claudia, Mike, Joe, Rob, Liz, Arun, Roger, Chris, Vlad, Harita, Valentina and Andy “King” Moore for all the discussions during this journey. As I had to travel extensively during my research, I also want to extend huge thanks to Michaela Scarle for dealing with all (sometimes-difficult) travel requests. You are a star!

A special thanks to Litsa Paraskeva and Håkan Sivencrona who have been pillars of support, and have provided me with guidance and an international dimension to this research. Thank you my friends.

I am ever so grateful to Professor David Mullins and Professor Steven Maggs for always being there for me and helping me see the bigger picture, not only in research but also for my career.

Penultimately, I want to thank Maa, Baba and Nani for their love and faith, and for the constant support. I hope I did you all proud and I know Dadu will be smiling from above.

Last but not the least; I want to thank my wife, Poonam, who has been by my side throughout this journey. Thank you for your understanding and the unending support and for all the sacrifices you have had to make to help me through this journey. I love you forever…

DECLARATION

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author. Parts of this thesis have been published by the author and a list of publications has been included in this thesis.

Siddartha Khastgir

INCLUSION OF PUBLISHED WORK

As part of the research presented in this thesis, ten publications were produced. Two of the publications were in peer-reviewed Q1 journals. Eight of the publications were in peer-reviewed conferences. A list of the publications is presented below along with a reference to the thesis chapters (and sections). For each publication, the author of this thesis was solely responsible for creating the knowledge and results and the named others provided comments only for modifications of the manuscript.

Peer Reviewed (Q1) Journals

1. S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, "Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles," Transp. Res. Part C Emerg. Technol., vol. 96, pp. 290–303, 2018.

Publication one captures the driving simulator study conducted to evaluate the effect of static knowledge on trust. The contents of this publication have been discussed in chapter five (section 5.3). The author designed and conducted the study. Other named authors reviewed the design prior to the study being carried out and provided comments to the manuscript. The author approved the final publication.

2. S. Khastgir, S. Birrell, G. Dhadyalla, H. Sivencrona, and P. Jennings, "Towards increased reliability by objectification of Hazard Analysis and Risk Assessment (HARA) of automated automotive systems," Saf. Sci., vol. 99, pp. 166–177, 2017.

Publication two captures the initial Objective HARA workshop conducted in USA. The contents of this publication have been discussed in chapter eight (section 8.4). The author designed and conducted the study. Other named authors provided comments on the manuscript. The author approved the final publication.

Peer Reviewed Conferences

3. S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, "Effect of Knowledge of Automation Capability on Trust and Workload in an Automated Vehicle: A Driving Simulator Study," Adv. Intell. Syst. Comput., vol. 786, pp. 410–420, 2019.

Publication three discusses the effect of static knowledge on workload which was captured as a part of the driving simulator study one (static knowledge and trust). The contents of this publication have been discussed in chapter five (section 5.3). The author designed and conducted the study. Other named authors reviewed the design prior to the study being carried out and provided comments to the manuscript. The author approved the final publication.

4. S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, "The Science of Testing: An Automotive Perspective," in SAE Technical Paper: 2018-01-1070, 2018.

Publication four introduces the concept of Hazard Based Testing which was proposed as a result of a semi-structured interview study. The contents of this publication have been discussed in chapter seven (section 7.1). The author designed and conducted the study using his industrial network. Other named authors reviewed the design prior to the study being carried out and provided comments to the manuscript. The author approved the final publication.

5. S. Khastgir, H. Sivencrona, G. Dhadyalla, P. Billing, S. Birrell, and P. Jennings, "Introducing ASIL inspired Dynamic Tactical Safety Decision Framework for Automated Vehicles," in Proc. of the IEEE Conference on Intelligent Transportation Systems, Proceedings (ITSC) 2017.

Publication five introduces the concept of Dynamic HARA. The contents of this publication have been briefly discussed in chapter eight (section 8.8). The author proposed the concept presented in this paper. Other named authors provided comments on the concept and to the manuscript. The author approved the final publication.

6. S. Khastgir, G. Dhadyalla, S. Birrell, S. Redmond, R. Addinall, and P. Jennings, "Test Scenario Generation for Driving Simulators Using Constrained Randomization Technique," in SAE Technical Paper: 2017-01-1672, 2017.

Publication six captures the difference in terminology for use case, test scenario and test case. The contents of this publication have been discussed in chapter six (section 6.1.1). The author proposed the concept presented in this paper. Other named authors provided comments on the concept and to the manuscript. The author approved the final publication.

7. S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, “Calibrating Trust to Increase the Use of Automated Systems in a Vehicle,” in Advances in Human Aspects of Transportation. Advances in Intelligent Systems and Computing, vol 484, N. Stanton, S. Landry, G. Di Bucchianico, and A. Vallicelli, Eds. Springer, Cham, 2017, pp. 535–546

Publication seven introduces the concept of stages of calibration of trust of a driver while using ADAS and ADS. The author proposed the concept presented in this paper. The contents of this publication have been discussed in chapter three and in Appendix one. Other named authors provided comments on the concept and to the manuscript. The author approved the final publication.

8. S. Khastgir, S. Birrell, G. Dhadyalla, D. Fulker, and P. Jennings, “A Drive-in, Driver-in-the-Loop Simulator for Testing Autonomous Vehicles,” Proc. Driv. Simul. Conf. Eur. 2015.

9. S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, “Identifying a gap in existing validation methodologies for intelligent automotive systems: Introducing the 3xD simulator,” in Proc. of the IEEE Intelligent Vehicles Symposium 2015, 2015, pp. 648–653.

10. S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, “Development of a drive-in driver-in-the-loop fully immersive driving simulator for virtual validation of automotive systems,” IEEE Veh. Technol. Conf., vol. 2015, 2015.

Publications 8 – 10 are on test methods for ADASs and ADSs and discuss the challenges associated while introducing the WMG 3xD Simulator for Intelligent Vehicles. The contents of these publications have been discussed in Appendix two. The author approved the final publications.

ABSTRACT

Automated Driving Systems (ADSs) offer many potential benefits like improved safety, reduced traffic congestion and lower emissions. However, such benefits can only be realised if drivers trust and make use of such systems. The two challenges explored in this thesis are: 1) How to increase trust in ADSs? 2) How to identify the test scenarios to establish the true capabilities and limitations of ADSs?

Firstly, drivers' trust needs to be calibrated to the "appropriate" level to prevent misuse (due to over trust) or disuse (due to under trust) of the system. In this research, a method to calibrate drivers' trust to the appropriate level has been created. This method involves providing knowledge of the capabilities and limitations of the ADSs to the driver.

However, there is a need to establish the capabilities and limitations of the ADSs which form the knowledge to be imparted to the driver. Therefore, the next research contribution lies in the development of a novel method to establish the knowledge of capabilities and limitations of ADSs (used to calibrate trust) in a reliable manner. This knowledge can be created by testing ADSs. However, in literature, an unanswered research question remains: How to identify test scenarios which highlight the limitations of ADSs? In order to identify such test scenarios, a novel hazard based testing approach to establish the capabilities and limitations of ADSs is presented by extending STPA (a hazard identification method) to create test scenarios. To ensure reliability of the hazard classification (and of the knowledge), the author created a novel objective approach for risk classification by creating a rule-set for risk ratings.

The contribution of this research lies in developing a method to increase trust in ADSs by creating reliable knowledge using hazard based testing approach which identifies how an ADS can fail.

ABBREVIATIONS

ACC	Adaptive Cruise Control
ADAS	Advanced Driver Assistance System
ADS	Automated Driving System
ANOVA	Analysis Of Variance
ASIL	Automotive Safety Integrity Level
AV	Autonomous Vehicle
BUC	Built Up-Cab
DDT	Dynamic Driving Task
DoAT	Degree of Appropriate Trust
DoIT	Degree of Incorrect Trust
ETA	Event Tree Analysis
FB	Feedback
FMEA	Failure Modes Effects Analysis
FMECA	Failure Modes Effects and Criticality Analysis
FTA	Fault Tree Analysis
HARA	Hazard Analysis and Risk Assessment
HAZOP	Hazard and Operability study
HBT	Hazard Based Testing
HFAC	Human Factors Analysis and Classification System
HMI	Human Machine-Interface
ISO	International Standards Organisation
LSAD	Low Speed Automated Driving

LSAV	Low Speed Automated Vehicle
NHTSA	National Highway Traffic Safety Administration
ODD	Operational Design Domain
OEDR	Object Event and Detection Response
RBT	Requirement Based Testing
RCA	Root Cause Analysis
SA	Situation Awareness
SAE	Society of Automotive Engineers
SD	Standard Deviation
SSQ	Simulator Sickness Questionnaire
STAMP	Systems Theoretic Accident Model and Processes
STPA	Systems Theoretic Process Analysis
SUT	Subject Under-Test
TCN	Trust Calibration Number
TOR	Take-Over Request
TTC	Time To Collision
UCA	Unsafe Control Action
UDP	User Datagram Protocol
VEHiL	VEhicle Hardware-in-Loop
ViL	Vehicle-in-Loop
VRU	Vulnerable Road User

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION TO THESIS

Chapter 1

Over the past five years, the automotive sector has seen a major evolution. New technologies, technology providers, start-ups (e.g. Tesla, Faraday Future, Byton etc.) have entered the automotive space. The beginning of the departure from the car ownership model with the introduction of the concept of Mobility as a Service (MaaS) is also a part of this ongoing change. The current transformation of the transportation sector has been aided in part by the introduction of driver assist technologies, i.e., Advanced Driver Assistance Systems (ADASs) and the rapid development of Automated Driving Systems (ADSs) by vehicle manufacturers, which remove the driver from the driving task in some situations. While early ADASs have been available on commercial models since late 1990s (e.g. Adaptive Cruise Control on Jaguar and BMW cars (Marsden et al., 2001)), ADASs have become ubiquitous more recently with the introduction of systems like Lane Keep Assist, Lane Departure Warning, Park Assist and Traffic Jam Assist.

The sudden push to introduce ADAS and ADS is in part led by some of the potential benefits that these vehicle automation technologies bring, e.g. increased safety, lower emissions, reduced traffic congestion, increased traffic throughput, decreased drivers' workload etc. The rapid introduction is also being helped by the rapid improvements in sensor technology and computation power.

While the introduction of automation has many potential benefits, it also introduces many potentially unsafe situations which may lead to accidents. In comparison to automation in other domains like aviation, nuclear, process, chemical etc., introduction of vehicle automation is a more recent phenomenon. When automation was first introduced in other domains, it was coupled with the occurrences of catastrophic accidents (e.g. Three Mile

Island disaster, China Air 006, Air France 447), some of which have repeated themselves over the years (Le Coze, 2013). Such incidents have led to an increasing emphasis on ensuring that the introduction of vehicle automation (ADAS and ADS) is done in a safe manner.

However, uptake of new technology by users is not only dependent on its technical features but also on how users perceive the introduced technology. ADASs have traditionally suffered low uptake by customers. Eichelberger and McCartt (2014) found that Volvo drivers used ADAS features Adaptive Cruise Control (ACC) and Lane Departure Warning Systems for only 51% of the time they spent on highways due to low trust in the systems leading to disuse. However, even more dangerous is the situation of misuse of ADAS and ADS due to over-trust. This was unfortunately seen in a fatal crash involving a Tesla Model S (NHTSA, 2017a). The driver had over-trust in the car's Autopilot feature and wasn't monitoring the road and a failure of the Autopilot led to a fatal collision with the rear end of a truck. However, another crucial factor, which added to the development of driver's over-trust in Autopilot, was the marketing of the system by the manufacturer and the use of the term "Auto-pilot" for an assistance feature. The term "Autopilot" gives an indication of the feature being autonomous when it was not.

In order to realise the benefits of vehicle automation, in addition to ensuring that the technology in itself is safe, it is important to ensure drivers (and users) of the technology trust and use such features in their everyday drive in a safe manner.

In this thesis, the author discusses how to ensure the use of ADAS and ADS in a safe manner, preventing any misuse of the systems. Misuse refers to the usage of the automated system in situations for which it has not been designed. Furthermore, the author also explores how to increase the use of ADAS and ADS, preventing disuse of the system. Disuse refers to the operator not engaging the automated system even if it is suitable for usage. Together, preventing misuse and disuse of the systems, will ensure "appropriate" use of the ADAS and ADS.

1.1. Aims of the research

Having briefly discussed the need to increase the use of vehicle automation while preventing misuse and disuse of automation to ensure that the benefits of automation are realised in a safe manner, the motivation for this research was: *to increase "appropriate" use of ADAS*

and ADS. Via an extensive review of literature, trust was found to be a key factor influencing the use of an automated system which led to the main aim of this research being:

"To increase trust in automation in vehicles."

Trust can be classified into two types: 1) trust in the system and 2) trust with the system. *"Trust in the system"* means the drivers' trust in the capabilities of the system and/or in the system's ability to do what it is supposed to do. *"Trust with the system"* means drivers' awareness or attitude towards the limitations of the systems and their subsequent ability to adapt their use of the system to accommodate for the limitations in order to realise the expected benefit from the system. Trust with the system is governed by accurate knowledge of the capabilities and limitations of the systems along with the real-time feedback about the working state of automation. Subsequently, the author set out to answer the following research question within the scope of this thesis:

"How to increase "trust in/with" automation in vehicles?"

The research question was driven by an aspiration to understand the development of trust in automation and ways of calibrating trust to prevent misuse or disuse of automation. In order to answer the central research question in a rigorous manner, the following research objectives were identified:

1. To develop a conceptual model for development of drivers' trust in automated driving systems

   One of the identified factors influencing drivers' trust is "knowledge" of a system's true capabilities and limitations. Based on how knowledge is imparted, knowledge can be classified into static knowledge, dynamic knowledge and internal mental model. Static knowledge refers to the understanding of the working of the automated system. Static knowledge is administered prior to the driving task and is akin to an owner's instruction manual, however with information at a higher abstraction level. Real time knowledge (or dynamic knowledge) refers to the real time information about the automated system (e.g. automation health, current state of the automation, near-future intentions of the automation). Internal mental model refers to understanding or prior beliefs influenced by external sources (e.g. word of mouth, media etc.). Since the two former types of knowledge can be imparted or established through system design, they were the focus of the research. This led to the formulation of the second research objective.

2. To evaluate the effect of knowledge (static and dynamic) on calibration of trust.

   Calibration of trust has been defined as "*the process of adjusting trust to correspond to an objective measure of trustworthiness"* (Muir, 1994). This research objective explores the use of knowledge (static and dynamic) in adjusting trust levels in ADASs and ADSs.

After having established that the knowledge about true capabilities and limitations of the systems could be used to calibrate trust to prevent misuse and disuse of automation, the next research question (answered in this thesis) was motivated by the desire to develop a method to create the knowledge which could be used to calibrate trust. Testing is a way to create the knowledge about the capabilities and limitations of ADSs. However, it is suggested that to prove ADSs are even 20% safer than human drivers, they need to be driven for more than 11 billion miles (Kalra and Paddock, 2016a), making it an unfeasible proposition. Thus an efficient identification of test scenarios to be used in the testing process was identified as a gap in literature, leading to the following research question:

"How to create test scenarios to establish the limitations of automated driving systems?"

To answer the research question, the following research objectives were identified:

3. To develop an understanding about the characteristics of a test scenario (especially for ADAS and ADS)

4. To develop a methodology for creating test scenarios (based on the identified characteristics)

Once the knowledge (to be imparted to the drivers) was created via testing, the next step involves the classification of knowledge (in terms of risk associated with it). The knowledge to be imparted to the driver includes both the failure and the risk associated with it. Automotive Hazard Analysis and Risk Assessment (HARA) comprises of severity, exposure and controllability ratings. A risk assessment is conducted by experts in the field of safety who have an in-depth understanding of the system workings and its failures, and is a specialised skill. However, any risk assessment method suffers from two types of reliability variations. Firstly, inter-rateability variation which is caused due to different mental models between different experts or different groups of experts. Experts' mental model refers to their beliefs about how they perceive severity, controllability and exposure. This belief will be influenced by not only their knowledge and experience in the industry but also the type of organisation and culture they belong to. Secondly, intra-rateability variation which is caused due to variation in mental models of the same expert or the same group of experts at

different points in time. However, the knowledge imparted to the driver/user of ADASs or ADSs needs to be done in a reliable manner which leads to the next research question:

How to improve the inter- and intra-rater-reliability of the automotive HARA process?

To answer the research question, the following research objectives were identified:

5. To develop a rule-set for conducting automotive HARA
6. To determine the ability of the developed rule-set for HARA in improving the reliability of the automotive HARA

1.2. Structure of the Thesis

This thesis is structured as follows (Figure 1.1):

- **Chapter 2:** Provides an introduction to Advanced Driver Assistance Systems (ADASs) and Automated Driving Systems (ADSs). While discussing various benefits, the author elaborates some of the challenges that need to be overcome to ensure that we realise the benefits of automation.
- **Chapter 3**: This chapter provides an in-depth discussion via a literature review especially focussed on development of trust in automation. Subsequently, one research question and two research objectives corresponding to it were identified. The review of literature on trust met research objective 1 (mentioned in section 1.1)
- **Chapter 4**: Provides a description of the research methodology being adopted to answer the research questions and to meet the research objectives identified in the literature, and discusses some of the common tools (driving simulator, questionnaires etc.) used in the various experimental studies.
- **Chapter 5**: Discusses research objective 2 and two driving simulator studies used to meet the research objective. The two studies establish the effect of knowledge (static and dynamic) of true capabilities and limitations of ADASs and ADSs on trust.
- **Chapter 6**: This chapter provides an in-depth discussion via a literature review especially focussed on testing (test scenarios and safety analysis) to create the knowledge used to calibrate trust. Subsequently, two research questions and four research objectives corresponding to them were identified.

- **Chapter 7:** Meets research objectives 3 and 4 on creation of test scenarios to establish the limitations of ADASs and ADSs. In chapter 5, once it was established that by providing knowledge, trust can be calibrated to the appropriate level, the knowledge needed to be created. This chapter discusses the proposed method to create knowledge by testing ADASs and ADSs, which can be used calibrate trust.
- **Chapter 8**: Discusses research objectives 5 and 6. One of the aspects of the knowledge creation process was to ensure that the knowledge about the risks communicated (to drivers) is generated reliably, giving identical results when used by different manufacturers. This chapter discusses a novel approach of rule-set based classification to increase the reliability of the created knowledge.
- **Chapter 9**: While each chapter has its own discussion / summary of the results, an overarching discussion has been presented in chapter 9 to link each aspect of the research and findings with the overall aim of the research. This chapter also discusses the merits and limitations of this research and possible future work that might follow the work already presented in this thesis.
- **Chapter 10**: Provides the final conclusions from the research work presented in this thesis.

Figure 1.2 maps the thesis chapters to the research objectives discussed in this chapter.

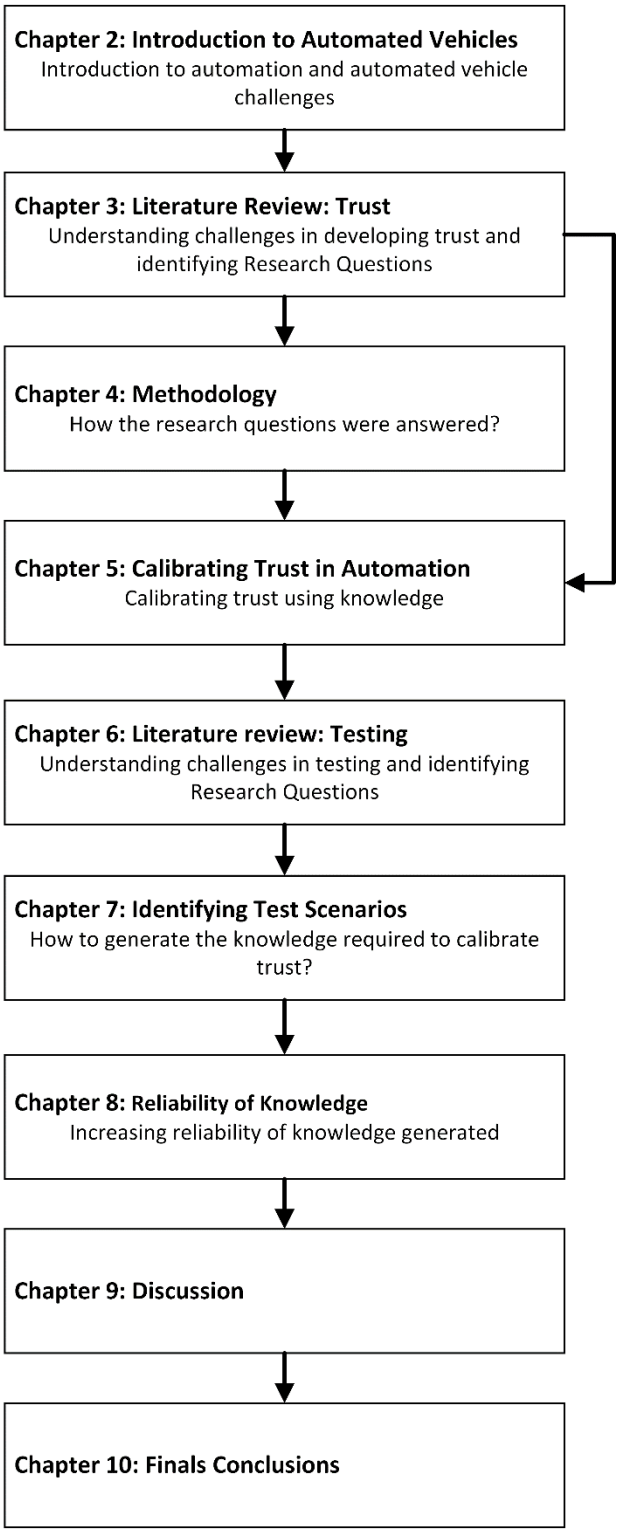

Figure 1.1: Structure of this Thesis' chapters

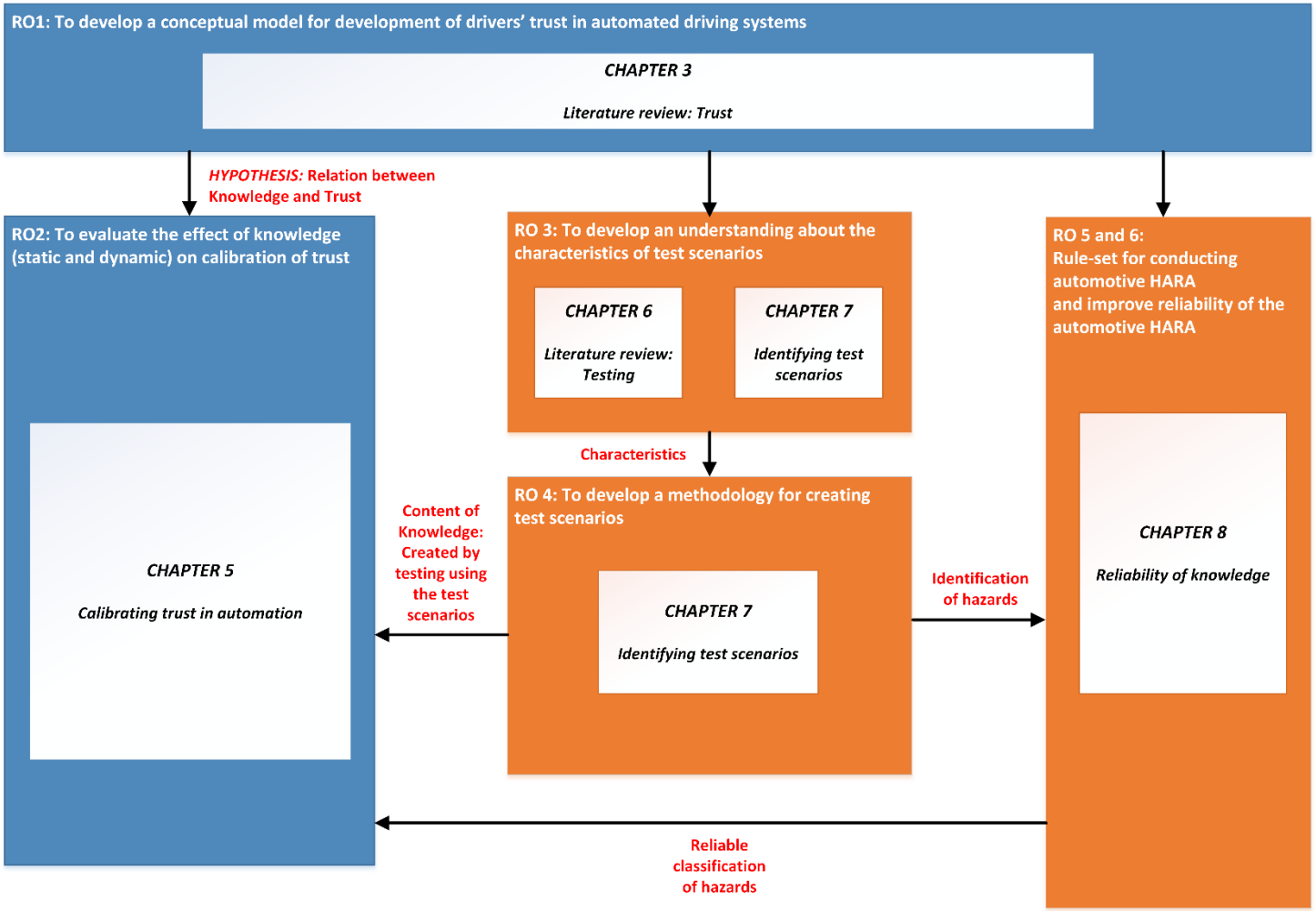

Figure 1.2: Thesis chapters mapped to Research Objectives

AUTOMATED VEHICLES: BENEFITS AND CHALLENGES

Chapter 2

An Introduction

In the last decade, there has been a gradual increase of Advanced Driver Assistance Systems (ADASs) (e.g. Adaptive Cruise Control (ACC), Lane-Keep Assist, Lane Departure Warning etc.) in on-road vehicles. More recently, there has been a push towards the introduction of higher levels of automation in vehicles with the aim of having Automated Driving Systems (ADSs). The push towards ADAS and ADS is driven by their many potential benefits like increased safety by reducing the number of accidents (Cicchino, 2017; Fagnant and Kockelman, 2015; Guériau et al., 2016; Tingvall, 1997), increased traffic throughput and road efficiency (Le et al., 2016; Talebpour and Mahmassani, 2016), time and monetary savings on parking (Fagnant and Kockelman, 2015), lower emissions (Fagnant and Kockelman, 2014), decreasing drivers' workload (Balfe et al., 2015; Stanton and Young, 1998) and providing more productive time to drivers (Cairns et al., 2014).

Additionally, an average customer is willing to pay more for ADAS and ADS (e.g. up to $3500 for partial automation and $4900 for full automation (Daziano et al., 2017)). While there is a variation among customers about their willingness to pay for automation, in a survey of over 5000 participants from 109 countries, it was found that 69% of the respondents believed that fully automated features will have 50% market share by 2050 (Kyriakidis et al., 2015).

It is evident from these statistics that the introduction of ADAS and ADS is inevitable and highly beneficial. However, to realise any of the many potential benefits of these technologies, it is essential that their introduction be done in a safe manner. Moreover,

Billings (1991b) suggested that introduction of automated systems changes the role of the human from an active manual engagement to a more passive monitoring task. Evidence from the aviation industry has shown that even though automation may lead to removal of some of the traditional manual errors, it can lead to the introduction of new errors (such as understanding of the automated system in order to be ready in case of a system failure). Therefore, the argument that autonomy itself will reduce the number of accidents is only partially correct.

This chapter discusses the various potential benefits that society at large could realise with the introduction of ADAS and ADS (section 2.2). Furthermore, the challenges that need to be overcome on the path to a safe introduction of automation in driving systems are also discussed (section 2.3). However, before discussing in detail the benefits and challenges of ADAS and ADS, it is important to understand the types of ADAS and ADS that exist and the basis for their classification.

2.1. Levels of automation

In order to better understand the types of automated systems, various organizations like National Highway Traffic Safety Administration (NHTSA), the German Federal Highway Research Institute (BASt) and SAE International have come up with their classification schemes for automated systems based on automation of driving functions and driver responsibility during an interaction with those functions. Winner (2016) proposed the triangle of autonomous driving with three corners of the triangle being 1) simple scenario automation (e.g. Adaptive Cruise Control, Lane Keep Assist) 2) low speed automation (e.g. Automated Valet Parking) 3) High risk automation (e.g. Collision Mitigation Braking, Emergency Steering Assist) and the goal being to travel towards the centre where all three meet . Michon (1985) suggested three levels of driving tasks: 1) Strategic 2) Tactical and 3) Operational. According to Michon (1985), strategic driving task involves *"general planning stage of a trip, including the determination of trip goals, route, and modal choice, plus an evaluation of the costs and risks involved."* Tactical driving task involves deciding the driving manoeuvres and avoiding obstacles in accordance with the strategic driving task. Operational driving task involves the execution of the tactical manoeuvres (e.g. steering and pedal inputs).

For discussion in this thesis, the classification scheme by SAE International has been adopted as it is the most widely accepted industry standard and is now being created in a joint SAE-ISO document also. SAE International in their standard J3016 (SAE International,

2018) have presented a 6 level classification from level 0-5. Figure 2.1 shows the classification scheme (as per SAE J3016) for the different levels and the distribution of responsibility between the driver and the system to perform the Dynamic Driving Task (DDT) and the DDT fallback task. Dynamic Driving Task is defined as *"All of the real-time operational and tactical functions required to operate a vehicle in on-road traffic, excluding the strategic functions such as trip scheduling and selection of destinations and waypoints, and including without limitation: 1) Lateral vehicle motion control via steering (operational); 2) Longitudinal vehicle motion control via acceleration and deceleration (operational); 3) Monitoring the driving environment via object and event detection, recognition, classification, and response preparation (operational and tactical); 4) Object and event response execution (operational and tactical); 5) Manoeuvre planning (tactical); and 6) Enhancing conspicuity via lighting, signalling and gesturing, etc. (tactical)"*(SAE International, 2018). It also uses the Operational Design Domain (ODD) as a parameter for the classification. ODD is defined as, *"operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not*

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

Figure 2.1: SAE Levels of Automation as per SAE J3016

limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics" (SAE International, 2018).

Furthermore, SAE J3016 defines DDT Fallback as *"The response by the user to either perform the DDT or achieve a minimal risk condition after occurrence of a DDT performance-relevant system failure(s) or upon operational design domain (ODD) exit, or the response by an ADS to achieve minimal risk condition".* As we move from level 0 (no automation) to level 5 (full automation), the automation capability gradually increases. ADASs can be classified anywhere between SAE level 1 to SAE level 2. Systems like Adaptive Cruise Control (ACC) which perform only longitudinal control are an example of SAE level 1 system, where as a Traffic Jam Assist (TJA) and Tesla's Autopilot system are examples of SAE Level 2 system. Currently there is no commercially available passenger car which can perform both DDT and OEDR (Object and Event Detection and Response) even in a limited ODD. SAE level 3 (conditional driving automation) is considered to be a step change in the automation capability of the systems. Future systems in SAE level 3, 4 and 5 will be called as ADSs.

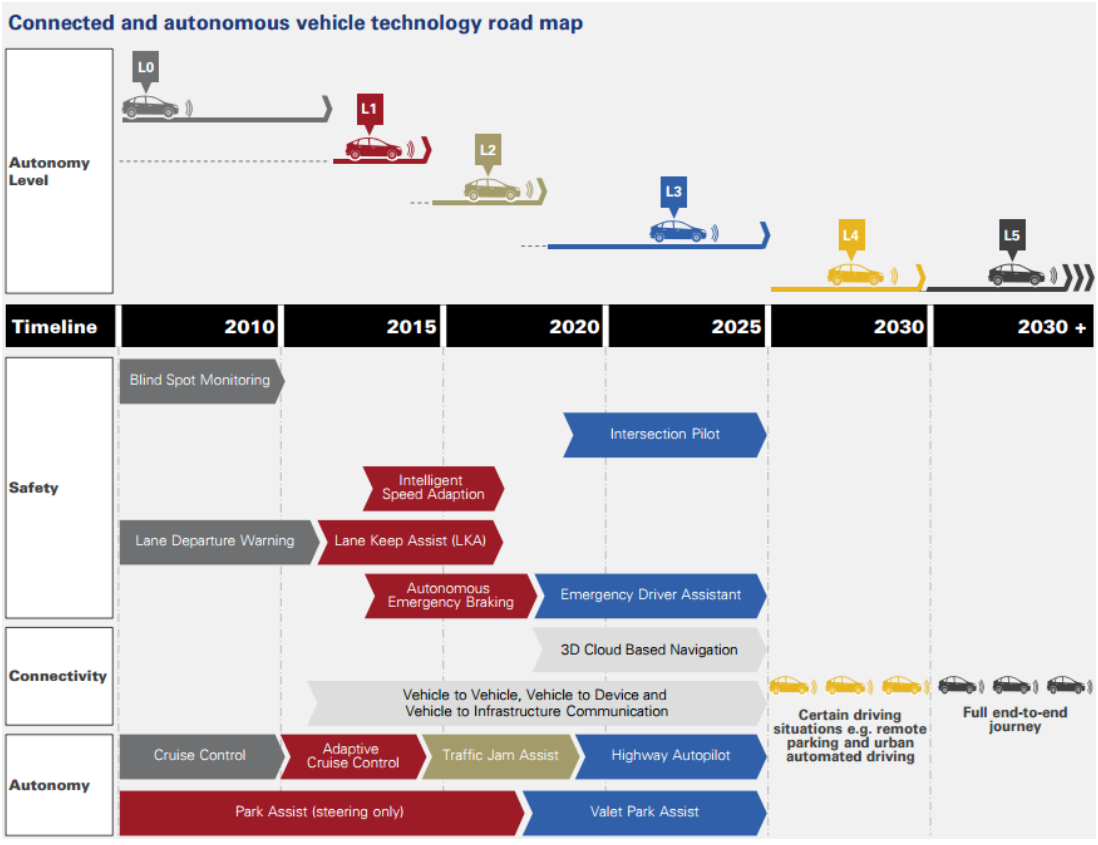

Figure 2.2: SAE Levels of automation, corresponding automotive systems and a potential introduction timeline (image source: KPMG: Connect and Autonomous Vehicles – The UK Economic Opportunity. March 2015)

With different levels of automation, different types of automated systems exist which provide varying degrees of assistance, from informative to warning to assistance. Drivers tend to use informative and warning systems more readily than assistive systems. This is because drivers are unwilling to give up control in critical scenarios. However, Parasuraman and Riley (1997) discussed that even an imperfect warning system (providing too many false warnings) can lead to disuse of the system.

While in SAE level 1 and 2, the driver is actively involved in the driving task, in SAE Level 3, the entire DDT (both vehicle motion and OEDR) is performed by the automated system, keeping the driver out of the loop. However, the driver is expected to perform the DDT fallback task in case of any failure of the system or in case the system has reached its operational boundaries (i.e., outside its ODD). The challenge here lies in ensuring that the driver has enough situation awareness about his role to perform the DDT fallback task and has enough time to perform the task. As we move to higher levels of automation (SAE Level 4 and 5), the automated system can perform both DDT and DDT fallback in limited and unlimited ODD respectively. Table 2.1 provides a summary of some of the ADAS systems already in market and the AD systems under development.

Table 2.1: Various ADAS and AD systems in market or in development (SMMT, 2019)

ADAS / ADS feature	**Level of automation**	**Commercial introduction**
Adaptive Cruise Control (ACC)	1	Available
Lane Keep Assist (LKA)	1	Available
Park assistance	2	Available
Traffic Jam Assist	2	Available
Tesla's Autopilot	2	Available
Audi's Traffic Jam Pilot	3	Potentially 2020-22
Highway pilot	4	In development (2025+)
Automated Valet Parking	4	2020 / 2021
Low Speed Automated Driving Systems	4	2022 - 2027
Robot Taxi	5	2035

2.2. Benefits

In 1996, Sweden adopted a "Vision Zero" policy which states that "eventually no one will be killed or seriously injured within the road transport system" (Tingvall, 1998). It brought together multiple stakeholders like vehicle manufacturers, road designers, state, city councils, municipalities and individuals, in order to achieve the mission of zero on-road

fatalities. According to Vision Zero's viewpoint, a holistic approach needs to be adopted. While changes in vehicles is a major aspect of the solution (with the introduction of passive safety, active safety and automated features), other aspects include changes in roads, streets, knowledge/awareness of individuals and legislations (Tingvall, 1998). While the principles of Vision Zero concept are valid for every country, the identification of changes and their implementation differs from country to country and the cultural aspect of the country needs to be taken into consideration in the strategic analysis plan (Johansson, 2009). With over 90% of the on-road accidents being attributed to human error (Singh, 2015), introduction of automated driving systems which replace the driver's driving tasks, have the potential to reduce the number of accidents and increase safety.

It has been suggested that the introduction of automated vehicles will increase road capacity and traffic efficiency by reducing the time-gap between vehicles (Bishop, 2000; Shladover, 2009; van Arem et al., 2005). However, Le Vine, Zolfaghari and Polak (2015) also suggest the existence of tension between the two factors: Increasing road capacity by reducing the time-gap between vehicles and driver comfort. Increasing road capacity might require acceleration and deceleration behaviour with which vehicle occupants may be uncomfortable. This would lead to a restriction in the allowable vehicle dynamics behaviour.

Automated driving systems can also reduce driver's stress and workload levels during driving by taking over the difficult and more mundane driving tasks (Balfe et al., 2015; Reimer et al., 2016) and subsequently increase a person's usable time (Cairns et al., 2014; Le et al., 2015).

Vehicles equipped with automated driving systems have the potential to increase usage efficiency of vehicles. Privately owned vehicles remain parked in garages or parking bays for over 95% of the time (Bates and Leibling, 2012). In case of automated vehicles, instead of being parked in the parking bays, these vehicles could be used to provide "transport services" like on-demand mobility, to increase their usage time. Not only will such services change the way cities are planned (e.g. fewer parking spaces), they will potentially change the car ownership model and reduce the total number of vehicles on road (Sparrow and Howard, 2017). Reduction in the total number of vehicles will also lead to reduction in vehicular emissions. An additional societal benefit of "transport services" inspired by automated driving systems is that they will provide mobility and independence for people with disability or elderly people who can no longer drive due to reduced eyesight.

2.3. Challenges

2.3.1. Reaping the safety benefits

While it is important to provide drivers the opportunity to use ADAS and ADS (with development in technology), it is equally important to ensure that the drivers use the systems in order to realise the potential benefits of such systems (Diels and Bos, 2015; Lee and See, 2004). Drivers' choice for use is influenced by their beliefs which form their internal mental model. Unfortunately, the usage of ADAS features like ACC and Lane Departure Warning has been traditionally low (51% of highway driving time (Eichelberger and McCartt, 2014)).

Exploring the possible variation in acceptance of Automated Vehicles (AVs) (which is influenced by user's propensity to trust AVs) due to culture and background, Haboucha et al. (2017) found that Israeli individuals are more willing to accept AVs as compared to North Americans, as the former care more for marginal cost as compared to capital cost. Similar results were found by Shin et al. (2015), where age of the respondents led to a heterogeneous mix of acceptance levels which suggested that early adopters of technology tend to be from the younger age group. This illustrates the influence of various factors on trust, acceptance and subsequently use of AVs. Interestingly, 44% of the respondents to Shin et al.'s survey mentioned that they would choose a regular vehicle over an AV.

Studies discussing the introduction of new technology in different domains like aviation, rail, automotive, etc. have shown that for the new technology to be accepted and used, effort needs to be made to introduce trust towards the new technology (Molesworth and Koo, 2016). Molesworth and Koo (2016) discussed that when participants were given a choice between conventionally piloted aircraft and remotely piloted aircraft (new technology), participants chose the former as they trusted it more, even though the latter was safer.

While the introduction of automated systems (e.g. ADAS or ADS) in the automotive industry is a relatively recent phenomenon, other industries like aviation, rail, process and manufacturing incorporated automation much earlier. Interestingly, while these industries have benefitted from automation, the introduction of automation in these industries was not smooth and faced various challenges, some of which continue to exist (Le Coze, 2013). Initially the introduction of automated systems suffered from reliability and safety issues. Gradually as these issues were eliminated via technological and process development advancements, a new type of issue emerged which was caused by human-automation interaction. Accidents like the Asiana 214 accident, China Air 006 accident (NTSB, 1986), the Three-Mile Island accident (Perrow, 1981), have occurred where the operator/pilot has incorrectly identified (due to over-trust) the state of the automation which led to the

accidents. The understanding of the human-automation interaction is still improving and unfortunately learning experiences have arisen from fatal accidents. Historically, similar accidents have repeated themselves in a 20-30 year cycle in different domains, referred to as the "déjà vu experience" (Le Coze, 2013) which emphasizes the need for a better understanding of the human-automation interaction in order to prevent similar accidents from recurring.

The potential for small and large scale fatalities is high in the medical, aviation, rail and nuclear industries. Thus, various procedures have been established to help identify the safety issues and reasons for the occurrence of any accidents (Olsen and Williamson, 2017). With the introduction of automation, a similar approach needs to be adopted in the automotive domain to identify system failures and driver-automation interaction failures to pre-empt the occurrence of a possible human error. Early detection of potential human errors and corresponding action can inhibit minor incidents from developing into catastrophic failures (Bennett, 2017).

Incorrectly designed automation can have negative impacts and thus, defeats the purpose of automation itself. Exploring this further, Bainbridge (1983) discussed the ironies of automation. Introduction of automation has been discussed widely among legislators and the automotive industry due to its ability to possibly remove the human driver from the control loop, thus reducing errors due to an unreliable and inefficient human driver. However, removal of the human driver makes the system prone to errors introduced in the system by the designer (another human influence) (Bainbridge, 1983), forming the first irony. Moreover, designers automate only the tasks they can automate without proper thought given to whether the task needs to be automated or not (Bainbridge, 1983). There is a likelihood that the designers' view of the best design of automation for a system and task distribution between the driver and the automated system differs from the driver's perception; and may potentially increase the workload of the driver due to the mismatch between driver's mental model and actual design (Sheridan, 1995). This forms the second irony. Thus, drivers' trust is affected which ultimately leads to misuse or disuse of the system if the system is not correctly designed or tested (Parasuraman and Riley, 1997). Misuse refers to the usage of the automated system in situations for which it has not been designed. Disuse refers to the user not engaging the automated system even if it is suitable for usage.

Once the designers have designed automation according to their understanding of requirements, the driver is left with the task of 1) monitoring the automated tasks 2) performing the non-automated tasks. A near-perfect automated system is more detrimental

for human-automation interaction than an unreliable automated system due to the rare need for human intervention in a near-perfect system. It leads to the human driver being out of the loop (not actively engaged in the driving tasks) for most periods. This means that the driver has to perform the monotonous task of monitoring the automated system, which as per Fitts list's (Fitts et al., 1951) is not a task humans are good at when compared to a machine. This has been said to be the third irony of automation (Bainbridge, 1983). Fitts suggested that tasks where machines outperform humans should be automated and tasks where humans outperform machines should be left to manual control. Thus, driving tasks that are automated need to be carefully selected and designed in order to enable drivers to use the systems in a correct and a safe manner. Therefore, designing automation in vehicles to ensure correct use of the ADAS and ADS remains a challenge.

2.3.2. Establishing the safety level

As improved safety has been advocated to being one of the major benefits of ADAS and ADS, it is important to ensure that the systems themselves are safe and reliable. However, both ADAS and ADS offer new challenges for testing and the safety analysis of the systems which are used to establish their safety level. A variety of ADAS and ADS exist or are in development, each of them offering a different kind of a challenge. As we move towards higher levels of automation in the SAE's levels of automation (level 0-5) (SAE International, 2018), the software and the electronic content in vehicles increase drastically. Over the past two decades, the software content in road vehicles has increased. It is said that cars have over 100 million lines of code compared to a jet fighter which has just over 1.7 million lines of code (Charette, 2009). Additionally, most of the upcoming innovations in the automotive industry have been credited to be due to software. With the increase in driver assistance and autonomous technologies, increased importance is being placed on rigorous Verification and Validation (V&V) processes which need to be established. While the ISO 26262 standard (ISO, 2018a), attempts to add clarity for V&V practices that should be followed for automotive software development to meet the functional safety goals, it falls short for higher levels of automation (Yu et al., 2016). This is mainly because for ADASs, the driver is always the fall-back, i.e., responsible for safe operation of the system. However, for higher levels of automation, the system must assume the fall-back role and bring the system to a safe state which increases the complexities and combinations of the scenarios in which the system needs to be tested.

In the absence of any common standards, there has been extensive discussion in the automotive industry regarding defining the completeness of testing for an automated system. Some experts have mentioned about 100,000 driven miles as a requirement for defining

completeness of testing, while others believe 11 billion miles are required (Kalra and Paddock, 2016b). However, the author is of the opinion that instead of the accruing miles, the scenarios experienced by the automated system during those miles are of more importance and should be the focus of research.

To meet the 11 billion mile challenge, role of simulation in testing of ADAS and ADS is gaining momentum (Bareket et al., 2003; Gordon et al., 2010), while many real-world Field Operational Test studies have also been conducted in US (Gordon et al., 2010; LeBlanc et al., 2006) and Europe (euroFOT project - (Benmimoun et al., 2012)). Various test methods like VEHiL (Vehicle Hardware-in-the-Loop) (Verburg et al., 2002), Vehicle-in-the-Loop (Bock et al., 2007), Co-ordinated automated driving (Schöner and Hurich, 2015) and test track testing have been developed for validation of ADAS and automated systems. Increasingly, driving simulators are also being used for testing ADAS and ADS as they offer a safer and a more reproducible environment for verifying and validating such systems. Real-world testing and test track testing (to some extent) produce the most valid results and can have the driver-in-the-loop for evaluation. However, the controllability and repeatability of scenarios is limited in such environments. On the other hand, VEHiL, Vehicle-in-the-Loop, Coordinated automated driving, are expensive setups and remove or modify the driver interaction with the system. While each of these test methods serves a specific purpose, they have a common requirement. All of these methods require the generation of test scenarios for which the systems are to be tested. An additional challenge with verification and validation of ADAS and automated driving is the need for testing for the huge sample space of system environment interactions. Without a robust verification and validation process, the chances of finding faults at later stages of development or even worse, after the release of the product increase. The increase of cost of bug fixing through a development cycle can be illustrated by the fact that it costs an average of $25 to fix a bug during development period which increases to $16000 after release (W.P. Klockwork, 2012). Any bug found after a vehicle's release is essentially a function of the number of affected vehicles. To fix the ignition switch issue affecting 2.6 million General Motors vehicles, GM had an estimated cost of $400 million (Malek, 2017). The identification of the *"unknown unknowns"*, (unforeseen events) which are also known as *"black swan"* scenarios remains a challenge for the research and the industry community.

2.3.3. Driver Take-Over Scenario

In SAE Level 3 automated system (discussed in section 2.1), the driver is not involved in the driving task for sustained periods, however is expected to take control and perform the DDT in case the automated system hands control back or in case of a failure of the ADS. When a

driver is not actively involved in the DDT, there is a tendency for the driver to be in an out-of-the-loop state. This is in line with the Fitts' list which suggested that humans are poor at monitoring systems as compared to machines (Fitts et al., 1951). Various studies (Dogan et al., 2017; Eriksson and Stanton, 2017a; Gold et al., 2013; Louw and Merat, 2017; Merat et al., 2012), have suggested that it takes anywhere between 5-40 seconds to bring the driver back-in-the-loop depending upon the driving situation, driver's state, vehicle velocity etc. However, no consensus has been achieved yet and more research is being carried out on take-over time.

Not only the duration of take-over time is important, the manner in which the driver is brought back in the loop is also under investigation. Different feedback mechanisms (audible, visual, haptic, multi-modal) are being evaluated to understand the optimum level and modality of feedback (Biondi et al., 2017).

2.3.4. Answering the ethical question

ADAS and ADS are capable of performing tactical and operational driving tasks. Thus, they are governed not only by technical requirements but social requirements also (Winner, 2016). Since an automated vehicle is essentially driven by software, the engineer or engineering team responsible for writing the software would be responsible for any accidents (ethically, if not legally). Such a situation brings to light the much debated "trolley problem" in an automotive perspective (Bonnefon et al., 2016; Keeling, 2019). Since the actions of the automated vehicle are essentially pre-coded by an engineer, how shall the engineer code the solution to the trolley problem?

Consider an example where an automated vehicle is going to have an imminent collision. However, in order to save the occupants of the vehicle, it can either swerve to the right and hit an old person or swerve to the left and hit a child (Bonnefon et al., 2016). The decision the automated vehicle will take has to be programmed by the engineers. However, any decision that the program makes will seem very much like a targeted decision. It is also suggested that trolley dilemmas for ADSs will not occur or are an incredibly rare phenomenon (De Freitas et al., 2019). With conflicting points of viewpoint, the trolley problem remains unanswered.

2.3.5. Legal and insurance considerations

One of the possible solutions to the trolley problem could be legislation. Legislation suggesting "minimal loss of life is the correct decision", could aid in the development of solutions for the trolley problem. However, like any new technology, ADSs necessitate new

policies and legislations. While incremental technologies can be catered by regular revisit of the policy framework, disruptive technologies (like automated vehicles) require a detailed review of the policies and in many cases may lead to a complete overhaul of the policy framework. One of the distinguishing features of disruptive technologies like automated vehicles is that not only the technology is radical, but also traditional vehicle manufacturers now face competition from non-traditional players like Google, Uber etc.

Due to lack of legislations, development of a new insurance framework for the situation where the driver is not driving, but the ADS caused an accident has been slow. The question remains: who is liable for the accident? A question which has stirred a lot of debate but still to yield any convincing answers.

2.3.6. Security and Privacy

It is in human nature to expect security and privacy in their lives. Therefore, any collection of data needs to adhere to the laws of privacy and security. Currently, humans knowingly or unknowingly share a lot of their data about their private and professional life. Features like location services on mobile phones, work and home details on navigation devices, e-commerce etc., collect all types of data. An automated vehicle will also have the capability to record large amounts of detailed data (vehicle parameters, GPS location, frequented destinations etc.) about the vehicle which will not only help accident investigations but also insurance providers. Most of the times, customers are unaware that their data is being collected by manufacturers or third parties (Markey, 2015). However, there are discussions about the ownership of the data and third-party companies trying to monetise the use of personal data coming from vehicles. Legislators are having extensive discussions on the data privacy laws and their enforcement process.

Additionally, the large amounts of data is vulnerable to a cyber-threat due to hacking. This may lead to individuals' identities been stolen by malicious attackers. Furthermore, the vehicle itself is susceptible to cyber-attacks (Markey, 2015). A connected vehicle extends the boundaries of the vehicle. It now interacts with other vehicles and infrastructure. In addition, many human-machine interface systems possess mobile connectivity too. All these external interfaces are potential attack points for cyber-attacks. In the widely discussed Jeep hack, Miller and Valasek (2015) demonstrated a cyber-attack on commercially available vehicle (Jeep Cherokee) running on the road and applying brakes on the car from a remote location. Subsequently, vehicle manufacturers, suppliers and data service providers have acknowledged the need to develop communication protocols for Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication to secure all such communications.

2.3.7. Sensor and control technology

Modern cars are equipped with variety of sensors (e.g. RADARs, ultrasonic sensors, cameras etc.) which are currently being used in SAE level 1-2 ADASs (e.g. Adaptive Cruise Control, Lane Keep Assist, park assist etc.). Future ADASs and ADSs will add even more sensors (e.g. LiDAR, high resolution cameras) which will enable the collection of detailed information about the vehicle's environment. However, adding more sensors to the vehicle increases the cost and complexity of the final product. Therefore, a trade-off decision between automation capability and cost (due to selection of sensor suite) needs to be made, while ensuring safe and optimum performance of the vehicle. In addition, the ability to achieve complete situational awareness based on the environmental data from various sensors is needed, especially in dynamic conditions (Winner, 2016). The current limitations on this aspect were highlighted in the fatal crash involving a Tesla Model S in which the Autopilot system used a combination of RADAR and camera (NHTSA, 2017a). While one of the sensors (camera) was able to detect the rear end of a white truck on a sunny day, the combined sensor suite and associated autonomous control algorithm chose to discount the sensor input, causing the fatal crash. Thus, not only is it important to ensure an optimum selection of the sensor suite, it is equally important to ensure that based on the data collected by various sensors, the vehicle is situationally aware about its environment.

2.3.8. In-Vehicle Design

As we move up the SAE levels of automation, the degree of automation increases and the drivers' role in the driving task decreases. In higher levels of automation (SAE level 3, 4 and SAE level 5) where the driver is not responsible for the Dynamic Driving Task (DDT) for long durations, the drivers (and passengers) may decide to do a variety of other (non-driving) tasks from reading newspapers/books, to watching movies, have official meetings etc. This shift in driving tasks has opened debates about how the interior of the vehicle should be designed to ensure the ideal user experience. However, while designing various in-vehicle layout concepts, an interesting issue of self-driving car-sickness has been identified (Diels and Bos, 2015). This is an interesting challenge as it is necessary to mitigate self-driving car-sickness in order to realise the benefits stated earlier, especially the idea of having more free-time for drivers / passengers.

2.4. Discussion

Like any new technology, ADAS and ADS have benefits and challenges associated with them (discussed in sections 2.2 and 2.3). As per the SAE Levels of Automation, ADASs are

classified between SAE Level 1-2. ADSs are classified between SAE Level 3 -5. ADAS systems (e.g. Adaptive Cruise Control, Lane Keep Assist etc.) have been commercially available in market since the late 1990s. One key and major difference between ADAS and ADS is the responsibility of the driver. In ADASs, the driver is always responsible for the safety of the vehicle. However, in ADSs the driver may be out of the driving loop for long sustained durations. Therefore, the safety of the system is a responsibility of the ADS (although in SAE level 3, the driver shall perform the DDT-fallback task). This change in driving role and responsibility makes some of the issues discussed in section 2.3 even more challenging. Thus, this thesis will focus primarily on ADSs and the challenges associated with commercial introduction and deployment of ADSs.

2.5. Summary

Automation in vehicles can take various forms from driver assistance to full autonomous control. In this chapter, the author introduces the concept of automation in the driving context and discusses the various potential benefits provided by the introduction of ADAS and ADS. These include:

- Increased safety
- Lower emissions
- Reduced traffic congestion
- Increased traffic throughput
- More free time available to drivers etc.

However, the author also discusses various challenges associated with the introduction of automation. Some of the main challenges include:

- Increasing use of available ADAS and ADS
- Ensuring the safety of ADAS and ADS
- Legal and insurance framework
- Security and Privacy while using ADAS and ADS

In order to realise any of the benefits mentioned earlier, it is important to find solutions to the challenges posed by the introduction of ADAS and ADS. In the next chapters (chapter 3 and chapter 6), some of the challenges (e.g. increasing use of available ADS and ensuring safety of ADS) are discussed in more detail and subsequently, research questions to be dealt within the scope of this thesis have been identified.

AUTOMATED VEHICLES: DEMYSTIFYING THE CHALLENGES - PART 1 (TRUST)[1]

Chapter 3

Literature Review

In chapter two, the author identified various opportunities and challenges associated with Advanced Driver Assistance System (ADAS) and Automated Driving System (ADS). However, in order to reap the benefits of the ADAS and ADS, it is essential that drivers use the systems. Therefore, it becomes imperative that solutions are found to the challenges offered by ADAS and ADS.

The two main challenges discussed and explored within the scope of this thesis are: 1) Trust and 2) Testing. In this chapter, the author discusses one of the challenges (trust) in greater detail and subsequently identifies associated research question which will be tackled within the scope of this thesis. While there is existing literature which attempts to find solutions to the challenges, this chapter and chapter 6, highlight the unexplored research areas and unanswered research questions.

The review of literature revealed that in order to increase drivers' use of the ADASs and ADSs, their trust in such systems needs to be increased. While exploring the concept of trust in the literature, the author proposes (hypothesises) that accurate knowledge about the capability of the ADAS and ADS can potentially increase trust. However, in order to generate this knowledge, ADAS and ADS need to be tested in a robust manner to establish

[1] Contents of this chapter have been published in the following publication:

Khastgir, S., Birrell, S., Dhadyalla, G. and Jennings, P. (2017) 'Calibrating Trust to Increase the Use of Automated Systems in a Vehicle', in Stanton, N. et al. (eds) Advances in Human Aspects of Transportation. Advances in Intelligent Systems and Computing. Springer, Cham, pp. 535–546.

their true capability and limitations and also the scenarios which illustrate their limitations need to be identified. In this chapter, the author discusses existing literature from various domains (aviation, process, chemical, nuclear and automotive) around trust. In chapter 6, the author discusses existing literature on testing (used to create the knowledge to increase trust) systems and the challenges associated with testing.

3.1. Introduction

In the USA over 90% of all on-road accidents are caused due to human error. Therefore, one may argue that the introduction of systems which assist or even replace the driver have a potential to reduce the number of accidents. This argument is based on the traditional belief as laid down by the Fitts' List (Fitts et al., 1951) which states that machines (automated systems) outperform humans in certain tasks. One such task is the ability of automated systems to react faster to a possible accident situation in comparison to an alert human driver (Carbaugh et al., 1998). Investigations of some of the major accidents in recent times in non-automotive domains have identified that human errors were amongst the causal factors (Lenné et al., 2012; Rasmussen, 1990; Shappell et al., 2007; D. Wiegmann and Shappell, 2001a) and have continued to be present despite the knowledge of various management methods (Salmon et al., 2010). In 2015, 1732 on-road fatalities were reported in the UK (Department for Transport HM Government, 2015), with the majority of those being due to driver error.

These statistics tend to suggest that human error is the cause of most accidents, however it is important to understand the nature of the human error and its causes. While the introduction of automation is a potential way to remove human error, incorporating automation also makes the system more complex and with it, the interaction between the human and the system also. It has been suggested that it is unfair to "criminalize" human error solely (Dekker, 2011) as "errors" in complex systems are inevitable and are caused by the complex interactions (Perrow, 2011). Therefore, automation, if not correctly designed into the system, can defeat the very aim with which it was introduced in the system, i.e., to reduce human error. However, if automation itself is inappropriately designed, it may be wrong to blame users or call it human error (Dekker, 2011). Before discussing automation further, it is important to understand "human error" and its types to understand design means to prevent the occurrence of human error.

3.2. Human Error

The changing nature of human-machine interaction and the introduction of more automation, has made the causal analysis of error challenging as the human task has shifted from an active manual role to a more passive supervisory role (Rasmussen, 1990, 1982). Skilled operators tend to adopt a dynamic approach to choose strategies to resolve conflicts in resource management to ensure performance levels and are driven by their knowledge base (Rasmussen, 1990). This dynamic approach adds to the challenges in the causal analysis of an error and also in the design of such systems.

Before the author delves into details of human error, it is important to define it. While it is hard to agree on a universally accepted definition for human error, this thesis adopts the definition of (human) error as: *"a generic term to encompass all those occasions in which a planned sequence of mental or physical activities fails to achieve its intended outcome, and when these failures cannot be attributed to the intervention of some chance agency"* (Reason, 1990). A series of planned activities may fail either if the execution of the activities didn't go as planned or if the plan itself was incorrect. This concept aids in classifying errors into two categories: slips (or lapses) and mistakes. Slips are *"errors which result from some failure in the execution and/or storage stage of an action sequence, regardless of whether or not the plan which guided them was adequate to achieve its objective"* (Reason, 1990). Mistakes are defined as *"deficiencies or failures in the judgemental and/or inferential processes involved in the selection of an objective or in the specification of the means to achieve it, irrespective of whether or not the actions directed by this decision-scheme run according to plan"* (Reason, 1990). Due to the subtle nature of mistakes, they are difficult to detect and hard to predict, whereas slips are more obvious in nature (Woods, 1984).

3.2.1. Types of Human Error (slips and mistakes)

Human errors tend to occur when the constraints designed in a system fail (Nancy G. Leveson, 2011). In order to understand the nature of errors, it is important to understand the human behaviour that may be adopted in realizing the designed constraints. Rasmussen introduced three different levels of human behaviour: skill, rule and knowledge based performance (Rasmussen, 1983). According to Rasmussen, Skill Based (SB) behaviour represents *"sensory – motor performance during acts or activities, …., take place without conscious control as smooth, automated, and highly integrated patterns of behaviour"*. An important aspect of SB behaviour is that it occurs intuitively (without conscious effort). Rule Based (RB) behaviour is *"controlled by a stored rule or procedure which may have been derived empirically during previous occasions"*. The distinguishing feature of RB behaviour

is that it occurs consciously with the human having the ability to report the rules used in performing the activity. However, when faced by an unfamiliar situation, humans need to resort to a higher conceptual level of behaviour: Knowledge Based (KB) behaviour. As the situations are unfamiliar, there are no rules or previously acquired skills available to deal with the situation. The behaviour becomes goal-controlled and requires an understanding (i.e. a mental model) of the structure of the system and surroundings to make a cognitive decision (Rasmussen, 1983). This distinguishes KB behaviour from RB and SB, and its significance in human-automation interaction will be discussed in section 3.5 and section 3.6.2. Inspired from Rasmussen's SRK framework, human errors can be similarly classified into three categories (Reason, 1990):

- Skill based slips
- Rule based mistakes
- Knowledge based mistakes

The three categories are distinguished by the fact that whether or not a human is involved in the process of problem solving at the time when an error occurs (Rasmussen, 1983; Reason, 1990). SB slips are precursors to the detection of a problem as they occur unconsciously. During RB and KB mistakes, the human is actively involved in finding a solution to the problem. According to Reason, SB slips tend to occur due to improper/insufficient attention, whereas RB mistakes occur due to the inability of the human to match the rule that is needed for that situation. It is important to know that the correct rule required for the situation is present in the person's mental model. KB mistakes are characterised by absence of any rule or experience as they occurs in an unfamiliar environment. While majority of errors which occur in daily life are either SB slips or RB mistakes, their probability of occurring is lower as compared to KB mistakes. SB is displayed primarily in routine tasks and non-challenging tasks. Challenging tasks require either RB behaviour (which means a stored rule-set is present in the mental model) or KB behaviour (in the absence of any stored rule-set). As humans have an extensive array of rule-sets that are ever evolving, the instances of display of KB behaviour are rare. In other words, KB mistakes have a higher opportunity cost. KB mistakes lead to decrease in trust, and subsequently lead to the disuse of the system. The author will discuss in detail how the opportunity cost for KB mistakes could be reduced in a driver-automation context in section 3.6.2.

As discussed in section 2.3.1, while an automated system removes human from the control loop, it makes the system prone to the errors introduced by designers (another human influence) (Bainbridge, 1983). For complex systems, there is the possibility the designer's choice for the best design for a system and the distribution of tasks between the human and

the machine, and the human driver's perception of the task distribution do not match (Rasmussen, 1990). This may ultimately affect a user's trust in the system (in case automation is not correctly designed) and his/her ability to use the system. Even though there is a mismatch in the perceptions or sometimes absence of operators' consideration in the design of the system, most accident investigations still blame the human operator as a cause for an accident and not a possible flawed design of the automated system or in other parts of the system or user training (Leveson, 2017). Additionally, introduction of automated systems introduces the possibility of new errors on the part of the human operator: 1) during monitoring the automated tasks 2) during performing the non-automated tasks. One way of compensating for the difference in the mental model of the human (driver) and designer, is by providing more skilled training to the human to better familiarize the human. While this approach is valid for aviation, chemical, nuclear etc. industries, this approach may not be feasible for the driving context due to the large number of vehicles and huge variation in driving abilities of humans (even though all drivers would have passed their driving test).

Introduction of automation in flights has been coupled with degradation of pilots' skills (Wiener and Curry, 1980). Long haul pilots tend to be more prone to skill degradation as compared to short haul pilots (Haslbeck and Zhang, 2017). In an aircraft simulator study involving 51 professional pilots, Haslbeck and Zhang found that the visual scanning pattern of long haul pilots was inappropriate for manual landing task and the need to address this skill degradation issue by providing specific training to the pilots in order to prevent Skill Based (SB) lapses. In sub-section 3.3, the author discusses the continuing occurrences of SB, RB and KB errors in different domains.

3.3. Types of "use" of automation

An operator's use of automation is influenced by many factors which will be discussed in detail in section 3.6. However, qualitatively, use of an automated system can be classified as (Parasuraman and Riley, 1997):

- *Misuse*: Misuse refers to the usage of the automated system in situations for which it has not been designed. This could potentially be a result of over-trust on the system. Instances of misuse have caused fatal accidents: e.g. China Air 006 (NTSB, 1986)
- *Disuse*: Disuse refers to the operator not engaging the automated system even if it is suitable for usage. This may be a result of lack of trust on the system. Disuse or distrust of automation has caused many accidents: for example Überlingen mid-air crash (BFU, 2004), Air France 447 (BEA, 2012)

- *Correct use*: Correct use of an automated system refers to the driver engaging the automated system only for the tasks and situations for which it has been designed. This requires the driver to have knowledge of the workings of the automated system, and acceptance of its limitations.

Table 3.1: Some accidents caused by Misuse and Disuse of automation

Type of use	Domain	Accident (year)	Type of Error	Brief Description
Misuse (Over-trust)	Aviation	Asiana 214 (2013)	SB slips	Pilots over-trusting the autopilot flight director system and auto-throttle system due to fatigue causing complacency and not monitoring the air speed gauge.
		TransAsia 235 (2015)	RB mistakes	Pilots didn't abort take-off when the ARM pushbutton didn't light. Pilots incorrectly identified the defective engine and shut down the operative engine. Pilots didn't respond to stall warnings.
		Kegworth air crash (1989)	RB mistakes	Pilots incorrectly identified the defective engine and shut down the perfectly working engine (engine 2). They didn't evaluate the reading on the engine instrumentation before shutting down engine 2.
		China Air 006 (1985)	RB mistakes	Pilots' over-reliance on the autopilot system after one of the engines lost thrust.
		XL Airways Flight 888 (2008)	SB slips	Crew carried out a demonstration to undertake a check of the low speed protections at a low height, while not taking into account the speeds mentioned in the programme.
		Tenerife (1977)	RB mistakes	Pilots incorrectly came to the conclusion that they were cleared for take-off.
	Nuclear / Chemical	Three Mile Island (1979)	RB mistakes	Operators didn't recognize the open state of the relief valve. This was due to lack of sufficient feedback from the instrumentation.
		Bhopal Gas (1985)	RB mistakes	Workers and plant managers didn't follow standard procedures as per plant rules
	Automotive	Tesla Autopilot	RB mistakes	Driver over trusted the Autopilot system and didn't monitor the environment and didn't keep hands on steering wheel.
Disuse	Aviation	Überlingen mid-air crash (2002)	RB mistakes	TU154M pilots distrusted the TCAS warning and over trusted the ATC instruction.
		Helios Air 552 (2005)	SB slips	Pilots didn't recognize the cabin warnings (cabin altitude warning horn, passenger oxygen mask deployment indicator and master caution)
		Air France 447 (2009)	RB mistakes	Pilots distrusting the stall warning and mistrusting (over-trusting) the flight director.

3.4. Automation in automotive context

The automotive industry needs to ensure that the introduction of ADAS and AD systems is not coupled with the introduction of human errors as experienced by other industries (aviation, chemical process, nuclear, petroleum etc.). Compared to other industries, introduction of automation within the driving context offers different challenges. Firstly, the number and density of vehicles on the road is far more than any other sector (aerospace, rail etc.). Secondly, each vehicle has its own driver, who may have a different behaviour and may interact differently with the vehicle. Thirdly, currently vehicle drivers don't undergo

any specialized training as pilots do in aviation, or operators do in manufacturing industries, making it even harder to bring all drivers to the same skill and understanding level for working with automated systems. While drivers also undergo a legally mandated driving test, the test doesn't address driving with ADAS or AD systems.

Automation in the driving context varies in the levels of autonomy and the role assumed by the driver (Banks et al., 2014). With different levels of automation, different types of automated systems exist which provide varying degrees of assistance, from informative to warnings, or assistance to full handover of control. In order to understand driver-automation interactions, the goal of the automated system needs to be put in context to the different driving tasks that are to be performed. Three levels of driving tasks are defined as 1) Strategic 2) Tactical and 3) Operational (Michon, 1985). In a manual drive, each of the three tasks is governed by the driver's mental model. This mental model is a result of the knowledge, rules and skills of the driver about working with the vehicle. The introduction of automation in vehicles has changed the driver's roles (in driving tasks) from being actively involved in the driving task to being a passive supervisor of the driving task (Merat et al., 2012). A fully automated system shall be able to perform all three driving tasks or both tactical and operational driving tasks with the human providing the strategic task input e.g. selection of destination and route driven. Between no automation and full automation, SAE J3016 (SAE International, 2018) defines different levels of automation. The responsibility of the different driving tasks changes in these levels. Table 3.2 depicts the division of driving tasks between the driver and the automated system for different automation levels. As a vehicle can potentially have SAE Levels 1-4 systems combined, the decision to shift authority between the driver and the automated system should be governed by the situation the vehicle is in and would differ from situation to situation.

This decision of handing over control between an automated system and the driver can vary in real time due to interactions between different factors like trust, knowledge, situation awareness, self-confidence and others (discussed in section 3.6). One of the fundamental bases for answering the question of division of driving task responsibility is to understand that technology development is just an enabler for the division of driving tasks and not the reason for task division. In other words, it is important to have an understanding of whether a system should be automated, rather than just knowing that a system can be automated, and which conditions can result in a successful driver-automation interaction (Wiener and Curry, 1980).

Table 3.2: Types of automation and its influence on task division

Type of Automation	Role of Driver	Role of Automated System
Informative and Warning (SAE Level 0)	All dynamic driving tasks. Perceive informative systems to increase situation awareness.	Only provide information or warning signals to possibly increase situation awareness.
Assistance (SAE Level 1/2)	Some dynamic driving tasks shared between driver and automated system. Object and event detection responsibility lies with the driver. Driver responsible for ensuring safety of the system and serves as fall-back.	Some dynamic driving tasks shared between driver and automated system.
Conditional Automation (fully automation in some scenarios) (SAE Level 3)	Passive role in some situations. Driver responsible for ensuring safety of the system and serves as fall-back.	Performing all dynamic driving tasks in some situations and providing feedback and increasing situation awareness to enable driver to perform the fall-back task. Object and event detection responsibility lies with the system.
High Automation (full automation in some scenarios with fall-back capability) (SAE Level 4)	None	Performing all dynamic driving tasks in some situations and has the ability to perform the fall-back manoeuvre to reach a minimal risk situation. Object and event detection responsibility lies with the system.
Full Automation (SAE Level 5)	None	All dynamic driving tasks. May provide SA to occupant.

In order to reap the benefits of automation, it is important to provide a system to the drivers which they would use. Design of an automated system cannot be analysed fully unless the interactions of the automated system with humans are considered in the design and analysis process (Karltun et al., 2017) in both adverse and ideal working environments (Sauer et al., 2013), as the interaction differs with changing environments. In addition, the human-automation interaction differs with age and gender too (Son et al., 2015). This emphasises that human drivers are not deterministic actors. They are driven by goals and actively search for new information in order to update their goals depending on the changing environments (Rasmussen, 1983). The dynamic nature of drivers' decision making process makes the driver – automation interaction a complex process which warrants more exploration. Therefore, factors influencing the use of ADAS and AD systems in a vehicle (if it were to be offered to the driver) need to be identified and will be discussed in section 3.6.

One of the key contributors to drivers' use of such systems is the drivers' trust on the systems (Lee and Moray, 1992; Lewandowsky et al., 2000; Muir and Moray, 1996). Many studies have discussed individual factors affecting trust and by way of theoretical and experimental results, have established a correlation between some of the factors and trust and use of automated systems (Inagaki, 2003; Lee and See, 2004; Mouloua et al., 2001; Rajaonah et al., 2008). However, a conceptual model of driver-automation interactions in an

automotive context has not yet been proposed. Thus, there is a gap in the existing literature to develop a conceptual model for the driver-automation interaction in the driving (automotive) context to aid the design of such systems to ensure their appropriate use. This gap has been addressed in section 3.6.

3.5. Drivers' "correct" use of automation

In a survey of Volvo drivers, Eichelberger and McCartt (2014) found that drivers used ADAS features like Adaptive Cruise Control (ACC) and Lane Departure Warning Systems for only 51% (of highway driving time) and 59% of all the time, respectively. Similarly, in another Field Operational Test study, over one-third of the drivers indicated that they would have turned the ADAS (FCW) off if they were given an opportunity (Ervin et al., 2005). While there has been low uptake of ADAS in general, there are certain examples of ADASs which have had a high uptake. For example, Honda's night vision system has received high uptake and high user reputation in Japan (Nakajima, 2008). This is due to the fact that the system's value in daily usage for customers is not restricted to rare occasions when the feature is invoked, e.g. accident scenarios. This strengthens the earlier argument that design of systems should be driven by usage value to the driver and not the technological ability to create a system. Technology itself should be treated as an enabler, not as the causal factor. The primary evaluation criterion for the benefit of automated systems should be functional performance in the context of human interaction with automated systems. In technology driven development of automated systems, the designers of the automated systems automate tasks they technically can automate, leaving the driver to perform the tasks which they can't automate, an irony of the introduction of automation (Bainbridge, 1983). While the move towards ADASs and ADSs is partly due to the increased technological ability to automate driving tasks (Hancock, 2014), the ADASs and ADSs have technical and safety boundaries which are defined by the designer and mostly not conveyed to the driver (Naujoks et al., 2015). Some of the early studies conducted using ADAS found that *"false alerts"* were a major concern for participants in a Field Operational Test study (Ervin et al., 2005; LeBlanc et al., 2006). However, participants were more accepting of the perceived benefits of the ADASs (Gordon et al., 2010). In the absence of this knowledge of the boundaries of safe operation, drivers tend to use the ADASs and ADSs in an inappropriate manner causing unsafe situations. This defeats one of the key purposes of the introduction of automation which is to improve safety, a trend initially also experienced in the aviation industry.

A near-perfect automated system is counter-productive to the intention of automation (Itoh et al., 2013). One possible explanation is that in a near-perfect automated system, the driving

task changes from being actively involved in driving to monitoring the automated system; a task in which human driver fairs worse as compared to a machine (Fitts et al., 1951). As per the Malleable Attentional Resources Theory (Young and Stanton, 2002), drivers' attention capacity decreases in low load conditions and increases their propensity to be distracted in low load conditions (Young et al., 2017). Thus, increasing their reaction times in critical situations (Young and Stanton, 2007). Therefore, it is suggested that the driver should periodically be brought back in the control loop from an out-of-the-loop situation to improve driver attentiveness while using automation (Casner et al., 2016). In a driving simulator study exploring this concept, Navarro *et al.* (2017) showed that partial automation or imperfect automation is better than a near-perfect automation with very few misses (i.e., automation failures). However, the study conducted used a lane departure warning system, which does not take over active control of the vehicle. Therefore, the results of the study need to be cautiously interpreted as an automated system which takes active control of the vehicle may not provide similar driver response. This difference in driver behaviour with automated systems that take active control is further illustrated in an experiment studying a semi-automated forward obstacle collision avoidance system, which found that the semi-automated system is not effective in avoiding collisions due to delayed or inappropriate action by the driver as the avoidance system was triggered by driver's intention (Itoh et al., 2013). The difference in results between an automotive context and other domains is apparent as a simulation study (automated cabin air management system) requiring the operator to take active control found contrary results to studies in driving context (Chavaillaz et al., 2016a).

Differing comprehension of working processes (between designer and user, i.e., driver) of automated systems can lead to incorrect use of these systems by the driver. Automation, if used incorrectly, can lead to catastrophic consequences. Many aircraft accidents have occurred as a result of pilots using the automated systems in an aircraft (e.g. autopilot) in situations for which they were not designed (Sparaco, 1995). Drivers engaging ACC without full knowledge of the system's true capabilities, use it in conditions like fog, rain, steep curves and forget to override the ACC system leading to accidents or near misses (Hoedemaeker, 2000). As described in section 3.3, incorrect use (misuse or disuse) of automated system has resulted in many fatal and severe accidents in other industries too.

In section 3.6, the author discusses approaches to ensure drivers' "correct" use of the automated systems by calibrating the drivers' trust on the automated system to the appropriate level as per the system capabilities and limitations.

3.6. Conceptual model for driver-automation interaction

One of the most widely used models for human-automation interaction developed from experiences in the aviation and manufacturing industry was proposed by V. Riley (Riley, 1996). When Riley's framework for the use of automated systems is applied in the automotive context, two aspects are inadequate. Firstly, interaction in an automotive context is fundamentally different from aviation and manufacturing (as discussed earlier). Secondly, while Riley's study consisted of a game-based experiment, it failed to capture an essential factor of "risk" or "felt-cost" for the user in the use of an automated system. In this section, the author presents a conceptual model for the development of trust on automated systems in the driving context and also for drivers' use of automated systems.

3.6.1. Trust

Trust is one of the most crucial factors for the driver to use the automated system (Lee and Moray, 1992; Lewandowsky et al., 2000; Muir and Moray, 1996; Parasuraman and Miller, 2004; Parasuraman and Riley, 1997; Walker et al., 2016). Drivers will give up control of a vehicle to an automated system, only if they trust the automated system (Muir, 1987). In an experiment where participants were given the choice to use different levels of automation, participants chose to use lower levels of automation for 95% of the time (Sauer et al., 2013). Participants chose higher manual control than to give-up control. An operator's use of the system has been shown to be correlated with the subjective measure of trust in an automated system or sub-system (Muir and Moray, 1996). Drivers do not use a well-designed and highly capable automated system, if they find it untrustworthy (Parasuraman and Miller, 2004). Ironically, misplaced trust (leading to misuse or disuse) in an automated system can lead to incorrect usage of the system (Parasuraman and Riley, 1997) with catastrophic consequences. For example Air France 447 crash (BEA, 2012), China Air 006 accident (NTSB, 1986), Überlingen mid-air collision (BFU, 2004), Three-Mile Island (Randolph et al., 1980) etc.

Before discussing details of the development of trust, it is important to define trust in driving context. In order to define trust, the author adapts the definition of trust from Lee and See (2004) as, *"a history dependent attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"*. The addition of the reference to *"history dependent"* is particularly important for this work because prior knowledge about the system's capabilities and limitations affects an individual's attitude towards a system, thus affecting their trust. Trust is said to be influenced by various factors (Lee and See, 2004; Walker et al., 2016; Xu et al., 2014). Additionally, author proposes that this can

also include knowledge, certification, situation awareness, workload, self-confidence, experience, consequence and willingness. The addition of the history dependent context is of key importance and will be discussed in section 3.7, where the concept of calibration of trust is introduced.

In order to reap the true benefits of an automated system, one needs the driver to develop *appropriate trust* to ensure *"correct use"* of the system (Lee and See, 2004; Parasuraman and Miller, 2004). Trust, like use of technology is influenced by many factors (Lee and See, 2004; Xu et al., 2014). Within scientific literature, trust is often discussed as a single construct. However, inspired by Rajaonah et al. (2008) who suggest two forms of trust: trust in automation and trust in the cooperation with automation; for the automotive context, the author classifies trust quantitatively into two forms:

- Trust ***in*** the system
- Trust ***with*** the system

"Trust in the system" means the drivers' trust in the capabilities of the system and/or in the system's ability to do what it is supposed to do. *"Trust with the system"* means drivers' awareness or attitude towards the limitations of the systems and their subsequent ability to adapt their use of the system to accommodate for the limitations in order to deliver the expected benefit from the system. Trust with the system implicitly means that the drivers are aware about the true capabilities, and limitations of the system, and are able to adapt their usage to overcome the limitations of the system in real-time. This paradigm of trust is going to be adopted in this thesis. Similar to the use of technology, trust is influenced by many factors (Lee and See, 2004; Xu et al., 2014). These factors will be discussed in greater details in the next sections.

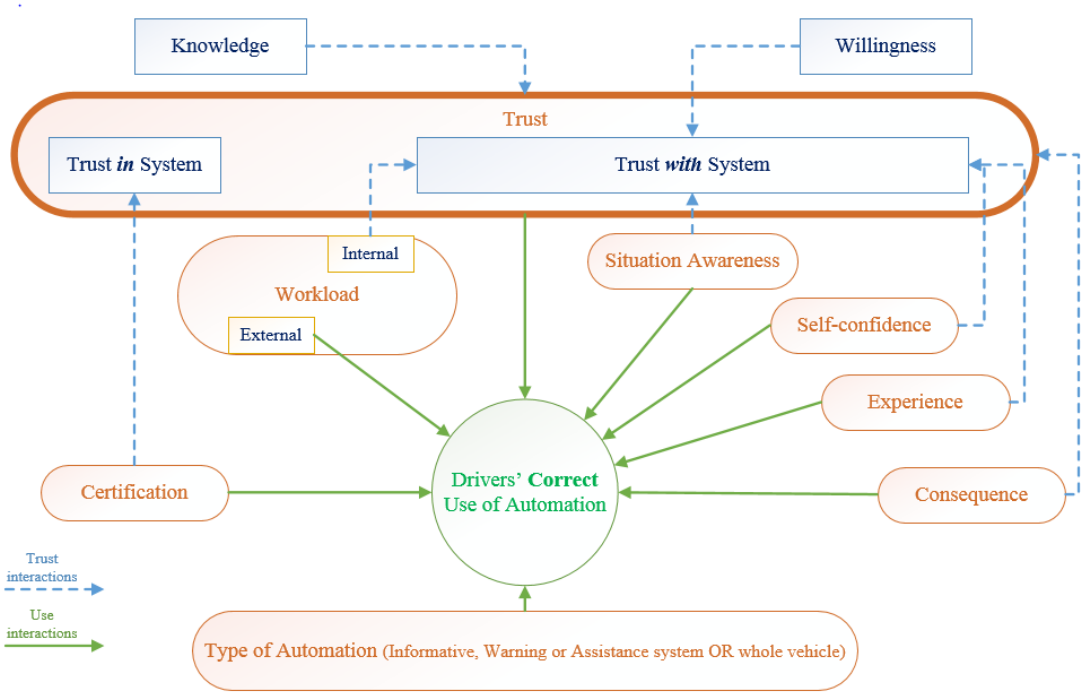

Figure 3.1: Framework for driver-automation interaction in automated systems in vehicles

3.6.2. Knowledge

In supervisory task domain, communication between the actors is key to development of trust (Ashleigh and Stanton, 2001; Stanton et al., 1997), by making the cooperation between the actors transparent (Casner et al., 2016; Nicola et al., 2013). In order to design a reliable system, the driver needs to be provided with the flexibility to cope up with an aberration rather than the constraints of an unknown system (Rasmussen, 1990). For a safe system, designers aim to have wide margins between normal operation and loss of control to enable easy detection of these boundaries by the driver. However, designers assume that the drivers will be aware about these margins and will be able to ensure reliable performance due to the flexibility being provided to them. This highlights an important requirement for driver-automation interaction: the need to make the driver aware about system boundaries for safe operation (both initially, i.e. statically and dynamically).

Knowledge about the automation capabilities and their current state encourage the development of trust. If the knowledge is factually correct and a true representation of the automation, it leads to the development of *"appropriate trust"*, leading to correct use of the system (by preventing disuse and misuse), and avoids introducing a false perception of capability, i.e., the driver having false expectations from the automated system or about system state (Sarter, 2008). The accuracy of the knowledge is especially important as it leads

to the formation of a mental model of the driver. A correct mental model of the system forms the basis of development of appropriate trust in/with the system (Kazi et al., 2007). Lenné et al. (2011) found in a driving simulator study that by providing training (imparting knowledge) to participants, appropriate use of the systems (larger following distance and reduced speed) could be induced. In a systems engineering world, where safety is considered a control problem, accidents are said to occur when the driver's understanding of the automation (driver's own understanding of the process model, i.e. driver's interpretation of the working of the automated system), is inconsistent with the system's actual (real time) process model (Leveson, 2017). Knowledge about the automated system is one of the factors that leads to the creation of driver's mental model (process model) of the system (Rasmussen, 1983). For appropriate use of the system and successful deployment, the driver's mental model needs to be aligned with system's capabilities (Reimer et al., 2016; Xiong et al., 2012). A survey of 130 participants who were owners of a Volvo XC60 with an adaptive cruise control system as a feature, (Larsson, 2012) found that 40% of the respondents lacked knowledge about the ACC system's (which is the simplest example of ADAS) limitations. Drivers need to understand the limitations of ACC systems, such as compromised performance around bends, or in adverse weather conditions, to safely use the systems (Rajaonah et al., 2006; Seppelt and Lee, 2007). A survey of 370 ACC owners, (Jenness et al., 2007) found that 72% of the customers were unaware of the warnings detailed by the manufacturer about the system's limitations.

With increasing automation, system complexity is bound to increase and drivers may be required to take-over in instances where the automation fails or is unable to deal with a situation. In such scenarios, drivers require accurate and comprehensible information about the system in order to appropriately supervise the system to ensure a safe take-over. Quality of supervision of an automated system by drivers is guided by drivers' understanding of how the system behaves in complex situations, system failures, along with the knowledge and understanding of their own role in terms of human actions to ensure safety (Jamson et al., 2013; Richards and Stedmon, 2016). Moreover, any failure occurring in an automated system does not negatively affect trust if the driver has the knowledge of the possibility of such failure(s) beforehand, i.e., is aware of the system limitations (Beggiato and Krems, 2013).

In the absence of knowledge of the system's capability, operators find it hard to calibrate their trust level according to reliability of the system causing lower level of trust in automations (Chavaillaz et al., 2016b; Lee and See, 2004). Based on literature (Bennett, 2017; Biassoni et al., 2016; Feldhütter et al., 2016; Miller et al., 2016; Rasmussen, 1985;

Seppelt and Lee, 2007; Xu et al., 2014), the following classification for knowledge about automated systems is proposed:

- Static knowledge: static knowledge refers to the understanding of the working of the automated system (Eichelberger and McCartt, 2014; Larsson et al., 2014)
- Real time knowledge: refers to the real time information about the state of the automated system (Banks and Stanton, 2015; Eriksson and Stanton, 2017b; Louw and Merat, 2017; Miller et al., 2016; Seppelt and Lee, 2007)
- Internal mental model: refers to understanding or influence of external sources (e.g. word of mouth, media etc.) (Biassoni et al., 2016), as drivers expect the ADASs and AD systems to behave as they were advertised (Casner et al., 2016).

Providing information about the true capabilities of an automated system leads to the development of "appropriate trust" (Dzindolet et al., 2003) and helps drivers to make rule-based and knowledge-based decisions. One of the key benefits of real-time knowledge is that it helps drivers make the correct knowledge-based decision by helping them select the correct rule from their rule set or adapt from it (Naujoks et al., 2014; Richards and Stedmon, 2016).

Louw and Merat (2017) demonstrated that by providing real time information about the reliability of an automated system, the driver could be brought back "in-to-the-loop", as it provides a link between the driver and the system (Banks and Stanton, 2015). However, the absence of static knowledge about the workings of the system leads to the occurrence of mode error and the inability of the driver to respond to a situation appropriately (Banks and Stanton, 2015; Sarter, 2008), which leads to lack of "trust in the system" and "trust with the system". Another dimension to knowledge is the accuracy of the information presented. If automation does not meet the initial expectations (information conveyed) about the automation capabilities, it can lead to disuse of automation and reduced "trust with the system". This is due to the larger cognitive effort required to recalibrate driver's mental model to have a correct representation of the system (Beggiato and Krems, 2013).

The classification of knowledge into the three types is discussed in detail in chapter five in which results from two experimental studies (conducted as part of this research) exploring the role of knowledge in development of trust have been included.

3.6.3. Certification

In order to define *"certification"*, the author adopts the definition from IEEE 24765 (IEEE, 2010) which defines *"certification"* as *"a written guarantee that a system or component*

complies with its specified requirements and is acceptable for operational use". By definition, certification establishes the safety level of a system in a binary manner, with certified systems being safe and uncertified systems being unsafe.

The existing literature fails to identify certification as an influencing factor for development of trust. The author's proposed conceptual model introduces certification as a factor influencing *"trust in the system"* and is the basis of the *"knowledge"* required to inform *"trust with the system"*. Thus, certification or testing to certify a system plays a binary role in the development of trust in an automated system and helps form the static knowledge that should be imparted to the driver to develop *"trust in the system"* and *"trust with the system"*.

A system which is certified "safe" and "tested" leads to the development of an inherent trust by the driver, i.e., the initial trust level. In a study involving 54 participants to evaluate the antecedents of trust in technology via a questionnaire study, Xu *et al.* (2014) found that competence of the automation is one of the main factors leading to the development of trust in the automated system. This is in line with results from earlier studies which stated that trust varies as a function of system's reliability (Muir and Moray, 1996).

In order to create this knowledge, it is the responsibility of the manufacturer/legislative bodies to ensure that a certified system does ensure the safety in all possible scenarios or explicitly mentions the conditions under which it has been tested (Schöner et al., 2009). Thus, there is a need for the development of test scenarios and test methods for these systems in critical conditions (Hendriks et al., 2010). Existing testing methodologies fail to test systems exhaustively, in order to mitigate the occurrence of acute failures and establishing system limitations (discussed in Appendix 2). If an automated system is able to perform safely only in some situations, the system should be certified for only those situations and the driver needs to be explicitly informed about the certified abilities and working situations. This will form part of the static knowledge imparted to the drivers and aid their knowledge based decision making leading to higher *"trust with the system"* and *"trust in the system"*. The knowledge of limited capability provides drivers with the ability to predict potential system failures or aberrations which helps them better prepare to mitigate or prevent accidents due to system failures, thus increasing their *"trust with the system"*.

In order to certify an automated system for all situations, it is important to identify all possible critical conditions by conducting an in-depth hazard analysis. Hazard analysis of automated systems with a driver-in-the-loop is still an open question which warrants further discussion. Thus, when a claim is being made for certification of an automated system, it is important to provide the context in which the certification has been made to provide accurate

information to the driver. The discussion about the method of knowledge creation or ways of testing and hazard identification have been discussed in chapter 6 and chapter 7.

3.6.3.1. Good failures and bad failures

In order to have a better understanding of the impact of certification, it is important to understand failures, their link with certification and subsequently drivers' trust on automated systems. Adapted from (Itoh et al., 1999)'s classification of failures, the author classifies failures into two groups: good failures and bad failures. *"Good"* failures are those which are predictable, repeatable and are known to the driver. Good failures may occur during the usage of the system. However, the predictable and repeatable nature of these failures will inform the driver not to use the system under the conditions which caused the failure.

"Bad" failures are those which are unpredictable, random, occur instantaneously and are hard to replicate. The aim of certification is to eliminate bad failures and identify good failures (i.e., system limitations) or to completely remove them by altering the systems in order to inform the driver of the system's true capabilities. The need for the development of use case scenarios for testing automated systems in critical conditions stems from this requirement. The spontaneous nature of bad failures makes it difficult for the driver to predict the occurrence of the failure leading to decreased *"trust with the system"*, whereas the knowledge of good failures aids in the development of *"trust in the system"* and *"trust with the system"*. Additionally, knowledge of good failures allows the driver to form rules regarding dealing with the failure, eventually leading to a rule-based behaviour. An additional dimension to *"good"* failures and *"bad"* failures is the temporal aspect associated with their occurrence. The prediction time associated with a failure will ensure whether the driver has the ability to intervene or otherwise control the situation. Thus, while good failures are predictable, they also provide sufficient time for the driver to react to them to ensure safe operation of the vehicle. On the contrary, bad failures are unpredictable and thus spontaneous in nature. Recurring bad failures can lead to disuse of the system (Parasuraman and Riley, 1997). The unpredictable nature of the bad failures forces drivers to resort to knowledge based behaviour which calls for high cognitive load which causes a decrease in the level of *"trust with the system"*. The challenges involved in identifying *"good failures"* to reduce the occurrence of *"bad failures"* is discussed in detail in chapter 6. Methods to overcome these challenges are discussed in chapter 7 and chapter 8.

3.6.4. Situation Awareness

Situation awareness (SA) is an emergent property of a system and a key factor influencing designer's design decision (Stanton et al., 2017). In the automotive context, SA is not only

crucial in helping the driver perform the new supervisory task due to introduction of automation, but also aids the ability of the driver to safely take back control when required (Beukel and Voort, 2017; Richards and Stedmon, 2016). SA affects both the development of trust and use of an automated system. With an increase in automation in systems, due to shifting of the driver's driving task, there is a possibility of loss of situation awareness of the driver (Mouloua et al., 2001). Irrespective of the complexity of task, humans are said to be poor monitors of automated systems (Parasuraman, 1987). Lack of situation awareness can lead to catastrophic results; Northwest Airlines MD-80 crashed at Detroit airport on take-off due to incorrect flap and slats settings. In its report, the National Transportation Safety Board (NTSB) mentioned the failure of the automated take-off configuration warning system as the reason for the crash and the reliance of the crew on it. They failed to realize that the aircraft had been incorrectly configured and when the automation failed, they were not aware of the state of the automated system and other critical parameters (NTSB, 1988). Inadequate information conveyed to the driver about the state of an automated system or over trust on automated system could therefore lead to complacency. SA is classified into three levels (Mouloua et al., 2001):

- Level 1 SA (SA1): perceiving critical factors in environment
- Level 2 SA (SA2): understanding the meaning of the perceived factors (individually and collective)
- Level 3 SA (SA3): predicting the future state of the automated system

SA is important to understand the distinction between the three levels and its influence on trust and use of automated systems. This differentiation holds more value when taken into account with the levels of automation in the automotive context. As the level of automation increases, the need for higher levels of situation awareness also increases. For a level 3 automated system, where the driver is supposed to take control in case of a system failure, the driver should have high level of level 3 SA and level 2 SA in order to anticipate a system failure and be ready for a manual takeover of the system. The ability to predict the future state of the automated system (level 3 SA) is developed as a result of the *"knowledge"* about the working of the automated system.

While automation may provide more situation awareness at one of the levels, it may reduce SA for other levels. This is true in a situation where the operator/driver is provided raw information from all sensors (thus increasing SA1), however, reducing their ability to comprehend the presented information (thus reducing SA2). Careful design and

consideration needs to be employed and a holistic approach needs to be adopted to increase situation awareness (SA1-3).

Situation awareness can vary with different levels of automation. A study by Endsley and Kiris (1995) found that situation awareness was reduced under fully and semi-automated systems than under manual control of an automobile navigation task. Interestingly, however, only SA2 (knowledge) was negatively impacted and SA1 remained unaffected. Although the drivers perceived the information (SA1), they were unable to understand it (SA2). Stanton et al. (2011) demonstrated that for a Stop & Go ACC system, by providing a representation of the radar system (representing SA2), drivers had higher detection rates for targets as compared to iconic display or flashing iconic display (representing SA1).

Norman (1990) suggested that feedback from automation is an aspect of keeping the operator engaged in the control loop. Automation of various tasks leads to new forms of feedback. However, the question remains unanswered on what is the correct form of feedback. For example, the introduction of electronic fly-by-wire flight controls in the F-16 fighter aircraft planes led to inappropriate control as the vibration which came through the flight stick was missing (even though the information was presented clearly on the visual displays). This led to the introduction of artificial "stick shakers" on the fly-by-wire systems to provide the appropriate feedback. Feedback such as situation awareness can be assumed to have different levels. Feedback of raw information can broadly lead to the development of SA1. Feedback from a "correctly" designed information display augmented by other forms (like auditory or haptic) can lead to correct knowledge and understanding of the information (i.e., development of SA2). The author proposes the following classification for feedback:

- Level 1 Feedback (FB1): feedback of raw information
- Level 2 Feedback (FB2): feedback of correct integrated (process) information

FB1 would increase SA1, while FB2 would increase both SA1 and SA2. While FB2 should be used most commonly for drivers, the author proposes that the drivers should have an ability to request for FB1 to aid their knowledge based behaviour in unfamiliar situations. SA2 is more important as drivers do not always possess the technical knowledge to understand the raw information and need to be presented with comprehensive processed information which may aid in their skill-based and rule-based behaviours. This differs from information presented to pilots; because pilots undergo appropriate training, they are better technically equipped to process raw information too. Thus, for aviation pilots both FB1and FB2 are important. "Correctly" designed information which may be provided to the driver by means of visual display or auditory message or haptic warning, can lead to correct

knowledge and understanding of the information (development of level 1 and level 2 SA). An in-vehicle speech assistant system (which provides process information to the driver) was found to increase situation awareness of drivers while positively affecting emotions too (Jeon et al., 2015). Type of feedback can be unimodal (Mohebbi et al., 2009) or multimodal (Biondi et al., 2017). Multi-modal feedback can be especially useful in conveying the urgency of the situation (hazard mapping) by manipulating the mode parameters, thus helping to achieve a cooperative teamwork between the driver and the automation (Dambock et al., 2013).

3.6.5. Self-Confidence

Various authors (Chavaillaz et al., 2016a; Lee and See, 2004; Lewandowsky et al., 2000; Riley, 1996) have discussed the role of self-confidence in trust and use of automated systems. Higher self-confidence is shown to cause lower use of automation (Riley, 1996). This suggests that high self-confidence has an influencing role on the decision to use automation. Furthermore, low self-confidence influences *"trust in"* the automated system. This can be explained as people having low self-confidence have more trust in the automated system's abilities than their own abilities (Lewandowsky et al., 2000). While Lee and See (2004) did suggest the relation between trust and self-confidence influencing usage, they failed to distinguish between effects of low self-confidence and high self-confidence. High self-confidence results in low *"trust in"* the system which may or may not result in decreased use of the system, depending on the interaction of trust with other factors.

3.6.6. Workload

While some studies suggest that a decrease in workload due to automation can lead to an increase in situation awareness (Billings, 1991b), others suggest a more independent relationship between workload and situation awareness (Riley, 1996). Endsley (1996) suggests that workload has a negative impact on situation awareness, but only at high levels of workload. Parasuraman et al. (1993) showed that with increases in workload, there is a possibility for an increase in the "complacency" or over-trust in the automated system.

Contrary to existing frameworks for driver's use of automated systems, and based on literature focussed on driver workload (Merat et al., 2012; Young and Stanton, 2002), the author classifies workload into two categories as both have differing effect on the development of *"trust with system"* and use of an automated system:

- *Internal workload*: this refers to the vigilance and mental effort involved in performing a task using automation and it affects "trust with the system". With increases in internal workload, "trust with the system" decreases.

- *External workload*: this refers to any secondary task a driver may perform (e.g. texting, reading etc.). With increases in external workload, tendency to use automation increases.

Monitoring of automated systems has been replete with issues of false alarms and warnings. This leads to the operator ignoring such alarms and reducing their level of *"trust in the system"*. Additionally, in the absence of knowledge, decreased automation reliability can also increase workload causing low levels of *"trust with the system"* (Chavaillaz et al., 2016b; Dadashi et al., 2013). This is due to the higher cognitive effort required to recalibrate the driver's mental model of the automated system to correspond to its true capabilities. Fuller (2005) suggested that driver task difficulty is inversely related to the difference between driver capability and driving task demand (i.e. workload). Collision or accidents tend to occur when task demands exceeds capability. Design of automation should thus aim to reduce workload while capability can be increased with experience or knowledge.

3.6.7. Experience

As discussed in section 3.6.1 and section 3.6.2, development of trust is a dynamic process. Trust in automation has been shown to increase with more experience with the automated systems (Beggiato et al., 2015; Chapman and Underwood, 2000; Hergeth et al., 2017). In a real-world study involving 15 participants, Beggiato *et al.* (2015) demonstrated that trust has a non-linear relationship with experience – specifically a power function. An experiment evaluating ACC usage by drivers, Pereira, Beggiato and Petzoldt (2015) found that the driver's use of ACC increased with experience, but only on motorways and not on urban roads. This illustrates that drivers' use and trust on automated systems is influenced by many factors. Irrespective of the age of the drivers, experience with automation helps build *"trust with automation"* and improves the quality of use of automated system. In a driving simulator study involving older (>=60years) and younger (<=28 year) drivers, Körber *et al.* (2016) demonstrated improved takeover quality with experience for both groups leading to less riskier manoeuvres between the first and the last situation. Similarly, Louw and Merat (2017) by measuring drivers' gaze dispersion, revealed that drivers' understanding of the system and hazardous situations increased with experience, leading to increased trust levels. It was found that after the first accident incident, the vertical gaze dispersion of the drivers reduced, indicating increased understanding of the system and possible hazards, and re-calibration of drivers' trust of the system. In a driving simulator study with 82 participants, Hergeth, Lorenz and Krems (2017) found that in a take-over scenario, trust in the automated system increased with experience in the second run. Experience helps calibrate the mental model of the driver to an appropriate and representative level leading to better driver performance (Lu et al., 2017). However, infrequent limitations or limitations which are not

experienced tend to “drop out” of the driver’s mental model. Thus, drivers need to be intelligently tutored to remind them about the system limitations to correctly shape their mental model (Beggiato et al., 2015). It has been demonstrated that with increasing experience, drivers’ workload decreases while using an automated system in complex situations (Paxion et al., 2015).

The author classifies experience into two categories: 1) experience with the system 2) experience of the external environment. While the former has been discussed in the preceding text, experience with the external environment can lead to complacency due to over-trust (Marieke H. Martens and Fox, 2007). In a real-world study involving 20 participants, Colonna *et al.* (2016) found that familiarity of external environment led to increased average speed among participants as compared to unfamiliar environments, thus leading to development of *“trust in the system”*. However, Martens and M. Fox (2007), and Martens and M. R. J. Fox (2007) found that with increase in (experience) familiarity of a route, driver’s glance duration at traffic sign decreases which led to less active information processing. This is due to the driver’s mental model which has resulted from experience. The pre-emptive nature of the mental model and decreased active information processing leads to losses in the situation awareness of the driver (SA2). When unexpected events occur, the mental model requires longer processing time, leading to higher cognitive loads to process the change in environment resulting in an inadequate or missing response.

3.6.8. Consequence

Automation of different types of tasks are affected differently by the presence and removal of the automation support (Cullen et al., 2013). Different tasks vary in their attributes (e.g. visual assistance, cognitive assistance, motor tasks etc.) and therefore the absence of automation has varied consequences. In a multiple task environment study, Cullen, Rogers and Fisk (2013) found that tasks which benefited the most due to the presence of automation (high criticality / low frequency tasks), were affected the most when automation was removed. It was found that a “felt-cost” or degree of consequence could affect the adoption of strategies and workload when dealing with complex situations (Paxion et al., 2015), with complex situations leading to higher workload levels. In an automotive context, it is hard for a driver to give-up control of the vehicle due the safety critical nature of the consequence. Thus, drivers are more accommodating of ADASs (assistance features) as compared to higher levels of automation. This may be because the degree of consequence is lower for assistance features as compared to higher levels of automation where they are expected to give-up control of the driving task for sustained periods of time.

3.7. Calibration of Trust

Calibration of trust has been defined as "*the process of adjusting trust to correspond to an objective measure of trustworthiness"* (Muir, 1994). Furthermore, calibration of trust is a result of dynamic interactions between the driver and the automated system and the driver can be shown to have characteristic stages. Initial development of trust on a system is easier to build than recovery of trust after a failure has occurred in the automated system (Lee and Moray, 1992; Muir, 1987).

Various stages of calibration of trust have been shown in Figure 3.2. However, it is important to note that the duration and rate of change of trust in each of the sections is a function of the type of automation, type of failure and the consequence of failure. The following stages characterize the process of calibration of trust on an automated system:

- Stage A: Initial phase: Influenced by static knowledge and an internal mental model of the driver
- Stage B: Loss phase: Influenced by failures (frequency and severity). The rate of decrease is influenced by the consequences of the failures
- Stage C: Distrust phase.
- Stage D and stage E: Recovery phase.

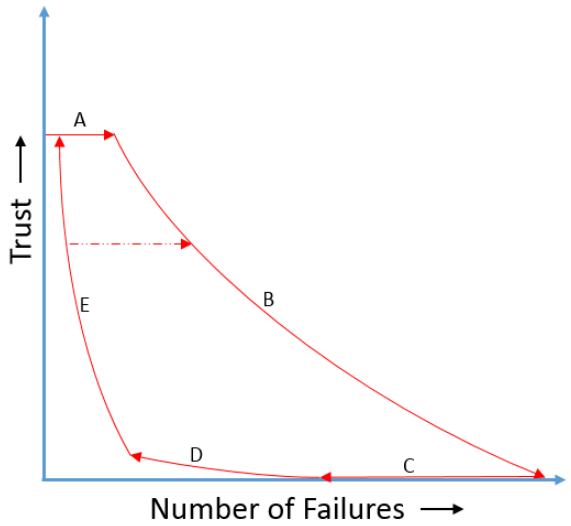

Figure 3.2: Trust calibration graph: Representing trust hysteresis

As discussed in section 3.6.7, the development of trust has a non-linear relationship with failures (Beggiato et al., 2015). However, the author suggests that the slope of the various stages is a design parameter and can be manipulated based on the factors discussed in

section 3.6. In order to provide guidance to designers, the author proposes the concept of *Trust Calibration Number* (TCN). TCN has three aspects: 1) Area enclosed by the trust calibration graph, referred to as Area of uncertainty 2) Slope of stage B 3) slope of stages C, D and E.

$$\begin{matrix} Trust \\ Calibration \\ Number \end{matrix} = \text{Area of uncertainty} + \text{Fcn}_1(\text{slope of stage B}) - \text{Fcn}_2(\text{slope of stages C,D, E})$$

$$\boldsymbol{Design\ Goal}: Minimize(Trust\ Calibration\ Number)$$

The area enclosed in the graph (Figure 3.2) is defined as the *area of uncertainty*. The smaller the area of uncertainty, the better the design of the automated system. A smaller area of uncertainty refers to fewer changes in trust levels and more predictable behaviour of the driver's trust level. However, as area of uncertainty is a relative concept, in order to ensure high use of the automated systems, the absolute trust should also be high. Design goals for a design should potentially be to have a small area of uncertainty along with ensuring high initial trust. A smaller area of uncertainty can be achieved by either an original design of the system or by intervention methods to calibrate trust to an appropriate level. Reduction of the area of uncertainty can be achieved by various techniques depending on the cause of the reduction in the level of trust. A manifestation of each of the factors discussed in the framework for driver's use of automated systems (Figure 3.1) can be used as an intervention method to calibrate level of trust to appropriate levels and prevent rapid degradation of trust. For example, informing the driver about the state/capability of the automated system in real time is one of the intervention methods.

However, there are additional dimensions to trust development and they focus on the rate of trust development (in the recovery phase) and the rate of trust degradation (in the loss phase); both factors together comprise the Trust Calibration Number. The second term comprising TCN focusses on the rate of trust loss on a system. This is important as it affects the duration of the system use and reduces the potential benefits the system could provide if used. System design or intervention techniques should not only stem the loss of trust, but also monitor the rate of trust loss. The third term in TCN focusses on the rate of trust recovery. The faster the rate of trust recovery, the faster the driver would start using the system and potentially reap the benefit(s) of the system. Overall, as a designer, the design goal should be to minimize the TCN. A lower TCN signifies minimal changes in the level of trust on the automated system which is an ideal scenario for the use of the automated system.

The use of knowledge as an intervention method in the process of calibration of trust to increase *"appropriate"* use of ADAS and ADS (by preventing disuse and misuse) is discussed in Appendix A1 via a case example.

3.8. Discussion: Identifying the research question

In the past 30-40 years, the introduction of automation in various domains has led to a change in the role of the human performing a task. Ironically, this was coupled with many accidents occurring due to inappropriate design of automation or human error due to the changing role of the human in the task. Unfortunately, due to our inability to learn from accidents (Christophe and Coze, 2013), there has been a repeat of similar accidents after 20-30 years. Therefore, the benefits realised from automation have come at the cost of human life. This highlights that designers and engineers still do not correctly understand the human-automation interaction process and the need for more research in order to ensure future automated systems introduced in different domains are safe, and free from similar accident situations.

Similar to other domains, automation in the automotive context offers many benefits, e.g. increased safety, reduced emissions, increased traffic throughput etc. However, to realise those benefits, drivers need to use the automated systems provided to them. While the introduction of automated systems in driving context has been quite recent, the industry has already seen two major issues: disuse (due to distrust or lack of trust) e.g. ACC in cars (Nakajima, 2008), and misuse (due to mistrust / over trust) e.g. the Tesla Autopilot fatal crash (NHTSA, 2017a). While discussing existing literature on use of automated systems (in sections 3.3 - 3.6), the author identified trust as a key factor influencing use. Therefore, the first research question to be discussed in this thesis is:

RESEARCH QUESTION 1

How to increase *"trust in/with"* automation in vehicles?

In order to facilitate the process of answering the research question, two research objectives were identified.

Research Objective One

"To develop a conceptual model for development of drivers' trust in automated driving systems"

In order to ensure that drivers safely use automated systems (i.e. prevent the occurrence of human error) and have the correct level of trust in them, it is important to understand the factors affecting the development of trust and driver-automation interaction. An understanding of human-automation interaction (discussed in section 3.6) formed the base for understanding the factors affecting the development of trust and developing a conceptual model for the process of trust development. Thus, the first objective was met with the formulation of the conceptual model for drivers' trust of automated systems in vehicles (presented in section 3.6) which is based on theoretical evidence. The model details various factors affecting the use of such systems and the development of trust in them. To the best of the author's knowledge, the conceptual model presented in this thesis is the first attempt to comprehensively capture various factors influencing development of trust in automated systems in a driving context.

As discussed in section 3.6.1, trust is one of the key factors influencing use of an automated system and in order to realise the benefits of automated systems, drivers need to use the automated systems. Having identified various factors affecting development of trust, the author then explored how to increase trust to the correct level (calibrate trust to the correct level) by using the factors as an intervention method. To manage the scope of the research, the author identified one of the factors as "knowledge", which is discussed within the scope of the research. Thus, this formed the basis of the second research objective.

Research Objective Two

"To evaluate the effect of knowledge (static and dynamic) on calibration of trust"

The author has suggested in section 3.8 that knowledge comprises of three factors: 1) static 2) dynamic 3) internal mental model. While the latter is influenced by society, the first two components can be explicitly imparted to drivers as a part of system design. Thus, the second research objective was conceived to evaluate the effect of use of knowledge (static and dynamic) to calibrate trust.

Static knowledge may be presented via drivers' manual or a pre-use script. Dynamic knowledge may be presented to the drivers via a human-machine interface display to inform the driver about the true capabilities and limitations of the system.

3.8.1. Next research steps

If knowledge (static or dynamic) is found to have a statistically significant increase in trust levels, then the next research step will involve exploring methods for creation of the knowledge required to calibrate trust. Review of methods for creation of knowledge and the associated research questions are identified in chapter 6.

3.9. Summary

While automated systems offer many benefits in non-automotive industries too where they have been introduced, e.g. aviation, nuclear, chemical process, railways etc., their introduction in these industries was coupled with many accidents, some of which have continued to repeat themselves. The introduction of automated systems in the automotive industry has had an exponential growth in last decade. Automation has potentially many benefits such as improved efficiency, reduced emissions and safety by changing the role of the conventional driver in the driving task. Ironically, the changing role of the driver during the driving task due to the introduction of automation has potential safety concerns if the systems are not used appropriately. While there is a need to ensure that the drivers use these systems, so that their potential benefits can be realized, it is more important to ensure that the drivers use the systems correctly and in a safe manner. According to the existing literature, trust is one of the key factors influencing drivers' use of systems.

In this chapter, the author has discussed existing literature on trust and factors influencing the development of trust. Moreover, the author has identified a gap in the literature on a possible intervention method to increase or calibrate trust to the appropriate level with the introduction of knowledge. The discussion on the means to create the knowledge reliably is presented in chapter 6 – 8, where creation of test scenarios and safety analysis have been discussed.

3.9.1. Research Questions and Research Objectives

Based on the review of the literature on trust in section 3.6, the following research questions (RQ) and their corresponding research objectives (RO) have been identified:

RQ 1. How to increase "trust in/with" automation in vehicles?

- *RO 1. To develop a conceptual model for development of drivers' trust in automated driving systems*
- *RO 2. To evaluate the effect of knowledge (static and dynamic) on calibration of trust*

In the next chapter (chapter 4), the author discusses the research methodology applied to answer the research questions and meet the research objectives identified in this chapter and in chapter one.

HOW TO MEET THE RESEARCH OBJECTIVES?

Chapter 4

The Methodology

In chapter three, the author had mentioned that two main challenges will be explored within the scope of this thesis. These being: 1) Trust and 2) Testing. Subsequently, the research methodology is also split into these two sections shown as Blue (Trust) and Orange (Testing) sections in Figure 4.1.

While the concept of *"informed safety"* (green box in Figure 4.1) will be introduced in chapter 6, in summary it means informing the driver (via static and dynamic knowledge) about the true capabilities and limitations of the ADAS or ADS, and the real-time state of the ADAS or ADS. Informed safety needs to be created in a manner that the driver (human) correctly understands it. It is essential to ensure that a high level knowledge of the system's working is conveyed and understood by the driver to ensure safe use of ADAS and ADS to maintain appropriate trust levels.

In this chapter, the author discusses the research methodology undertaken to create informed safety and investigate its effect on drivers' trust in automated driving systems.

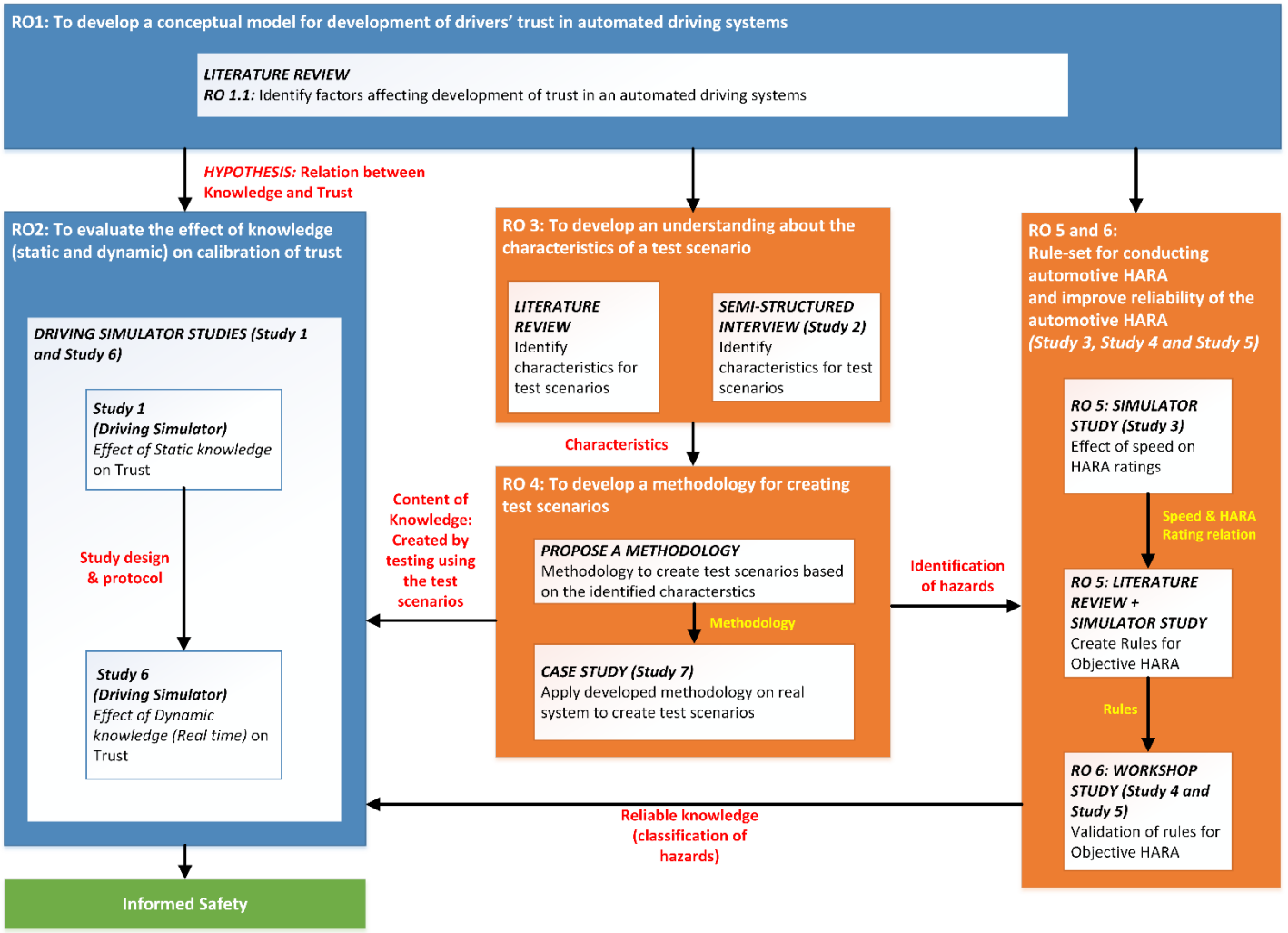

Figure 4.1: Schematic representation of research methodology and research stages mapped to various research objectives

4.1. Research methodology

The research methodology used to answer the three research questions (mentioned in chapter 1) have been discussed in this chapter. In order to answer research question 1, two research objectives (RO 1 and RO 2) were created. Meeting research objective one led to the creation of a conceptual model for development of trust. Research objective two focussed on one factor (knowledge of true capabilities and limitations of ADAS and ADS) of the conceptual model and explored its effect on trust and led to the creation of the concept of *"informed safety"*.

Research questions 2 and 3 were focussed on the creation of the content of knowledge (to be imparted to the drivers) in a reliable manner, with each of them being associated with two research objectives (RO 3-4 and RO 5-6 respectively).

For RO 2, RO 3, RO 5 and RO 6, both qualitative and quantitative methods were used to evaluate the subjective response. For RO 4, qualitative methods were used. Figure 4.1 shows a schematic representation of the research methodology and the various research stages which are mapped to the individual research objectives.

4.1.1. Methods for Research Question 1 (RQ 1)

"How to increase "trust in/with" automation in vehicles?"

In order to answer RQ 1, the author deployed a hybrid approach of literature review and driving simulator studies. Furthermore, two research objectives were created to answer research question 1 (as discussed in chapter 3) (blue section in Figure 4.1).

4.1.1.1. **RO 1:** *To develop a conceptual model for development of drivers' trust in automated driving systems* ***(via literature review)***

In order to develop a conceptual model for the development of trust in automated driving systems, the author needed to identify the factors that influence development of trust. To identify various factors, the author conducted a review of literature. The literature review encompassed literature from the aviation, process industry and the automotive domains. It was found that while extensive work had been done on the subject of human-automation interaction in the aviation and process industries domains, there was a limited understanding for the subject in the automotive domain. Additionally, various studies identified different aspects of human-automation interaction. Drawing inspiration from systems thinking, the intertwined nature of the relationships between the factors influencing human-automation interaction makes it essential to understand the concept as a whole (as a single system),

rather than evaluating individual factors independently. While effect of individual factors might be well understood separately, the interactions between factors could potentially display a different behaviour on development of trust. Thus, it was essential to develop a comprehensive conceptual model to understand the interactions between individual factors and trust on automated driving systems.

Once factors were identified (as a result of the literature review in chapter 3), a conceptual model (Figure 4.2) was proposed by the author to understand the development of trust in the driving domain. The author's focus in this thesis will be on one of the factors (from the proposed framework) influencing trust: knowledge.

"Informed safety" (mentioned earlier and discussed in detail in chapter 6) comprises of both "static knowledge" and "dynamic (real-time) knowledge", giving the driver the knowledge of the true capabilities and limitations of the systems. Based on the framework for development of trust, the author proposed the following hypothesis:

"Trust increases with accurate "informed safety" (provided in a static and/or a dynamic manner)."

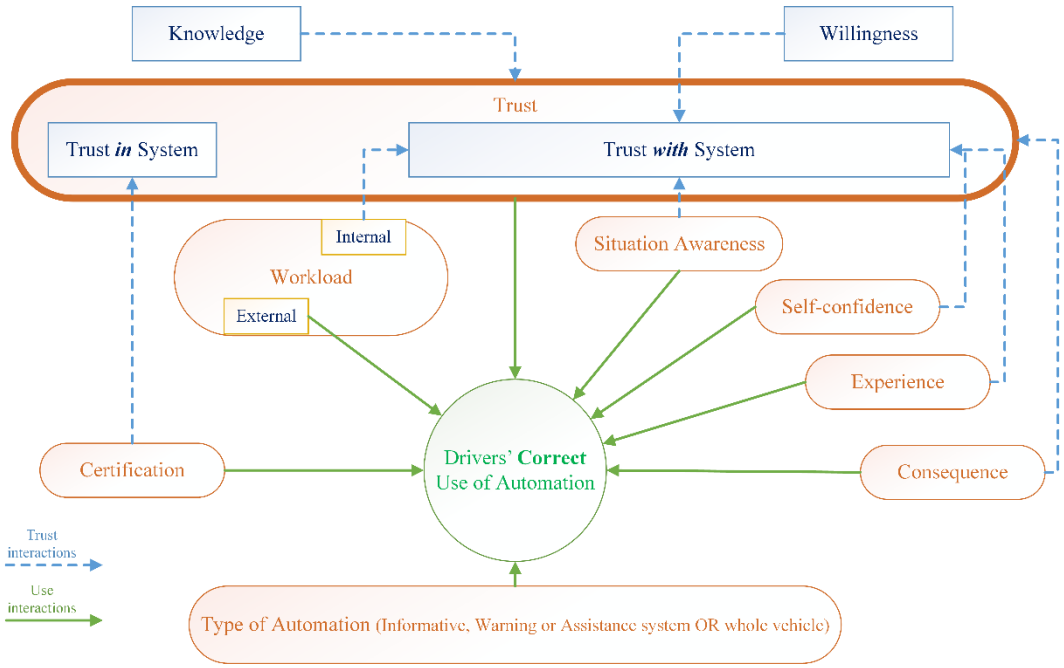

Figure 4.2: Framework for development of trust

4.1.1.2. RO 2: *To evaluate the effect of knowledge (static and dynamic) on calibration of trust* ***(via driving simulator studies)***

In order to evaluate the hypothesis as stated in RO 1 (and also answer the research question 1), RO 2 was designed and driving simulator studies were conducted to determine the relationship between static and dynamic knowledge with trust. Two separate studies were designed and conducted. **Study one** was designed to evaluate the relationship between static knowledge and trust. A driving simulator study involving 56 participants was conducted. **Study six** was designed to evaluate the relationship between dynamic knowledge and trust involving a 37 participants driving simulator study. In both the studies, participants experienced two different automated systems with contrasting capabilities. The studies were designed to evaluate the effect of *"informed safety"* on trust for two different systems (low-capability and high-capability automation). Some tools and methodological aspects of both the studies were similar and are discussed in section 4.2 and section 4.3, while the unique aspects of each of the studies are discussed in chapter 6.

One of the requirements for providing knowledge of the true capabilities and limitations of the ADS to the drivers leading to informed safety, was to create the knowledge in a reliable manner. To meet this requirement, research question two and research question three were created (discussed in chapter 6).

4.1.2. Methods for Research Question 2 (RQ 2)

"How to create test scenarios to establish the limitations of the automated driving systems?"

In order to answer RQ 2, two research objectives (RO3 and RO4) were created:

4.1.2.1. RO 3: *To develop an understanding about the characteristics of a test scenario (for ADAS and ADS) (via literature review and semi-structure interview study)*

In order to develop test scenarios for ADAS and ADS, an understanding of the characteristics of a test scenario needed to be developed. This was done via a literature review of test scenario generation in different domains (like aviation, computer science and automotive) and by conducting a semi-structured interview study (**study two**) involving verification and validation experts in the automotive domain. Participants were recruited from UK, Germany, India, Sweden, and USA and across the automotive supply chain, in order to avoid biasing the data due to nature of participants' organisations (OEM, tier-1 supplier, academia etc.).

Semi-structured interviews were adopted as they provide the flexibility to the interviewee to provide wider information and thus richer data, by enabling the formation of an

understanding between the interviewer and interviewee (Louise Barriball and While, 1994). Themes in the interview transcripts were identified via a coding analysis.

4.1.2.2. RO 4: *To develop a methodology for creating test scenarios (via case study)*

Scenarios present possible ways in which a system may be used to accomplish a desired function. After identifying the characteristics of test scenarios needed for ADAS and ADS, the author created a methodology to generate test scenarios. From the analysis of the semi-structured interviews, it was concluded that for ADASs and ADSs, it is essential to test *"how a system will fail"* in addition to *"how a system works"*. Thus, based on the analysis of the semi-structured interview study, the author developed a two branched approach to test scenario definition: hazard based test scenario definition and requirements based test scenario definition. Requirements based test scenario definition has existed in the automotive industry and is a standard practice. In this thesis, the author focussed on hazard based test scenario definition or hazard based testing.

Hazard based testing comprises of three steps: identification of hazards, creating test scenarios from hazards and classifying risks associated with the hazards reliably. For the first step (hazard identification), the author proposed the use of Systems Theoretic Process Analysis (STPA) as it captures the analysis of system interactions in the most efficient manner as compared to other hazard identification methods. For the second step, the author created an extension to the STPA process to create test scenarios. This (extension) process involves parametrizing elements of the STPA process outputs to create test scenarios. One of the features of the method created, is the ability to create both test scenarios and the pass/fail criteria for the test scenarios. To demonstrate the application of the created method, it was applied to a real-world system (low-speed automated driving system).

To perform the third step of hazard based testing, i.e., classifying risk associated with the hazards reliably; led to the formulation of research question three.

4.1.3. Methods for Research Question 3 (RQ 3)

*"**How to improve the inter- and intra-rater-reliability of the automotive HARA process?**"*

In order to impart knowledge of "informed safety" and its classification, there is a need to ensure that the "informed safety" concept is reliable. Reliability refers to the *"extent to which a framework, experiment, test, or measuring instrument yields the same results over repeated trials"* (Pg.11) (Carmines and Zeller, 1979). This suggests that the process of creation of the knowledge (necessary to create informed safety), i.e., hazard identification and risk classification and the manner in which the systems are tested, needs to be reliable.

Since the hazard identification and test scenario method have been discussed in RQ2, reliability of the risk classification process which forms part of the dynamic knowledge is explored in this research objective.

At the onset this seems to be a trivial problem, however the author demonstrates via literature review (from non-automotive domains) and conducted studies, that even experts in the field vary in their understanding and implementation which may lead to unreliable results for hazard classification in the automotive domain. To address this unreliability in the automotive Hazard Classification and Risk Assessment (HARA), the author created a rule-set for objectively performing HARA, i.e., classifying hazards for automated driving systems. To answer research question three, two research objectives (RO 5 and RO 6) were created:

4.1.3.1. RO 5: *To develop a rule-set for conducting automotive HARA* (via literature review, driving simulator study and feedback from international functional safety experts)

To develop a rule-set for automotive HARA, the author parametrized the three different aspects of an automotive HARA (severity, exposure and controllability ratings).

Controllability is defined as *"an estimate of the probability that the driver or other persons potentially at risk are able to gain sufficient control of the hazardous event, such that they are able to avoid the specific harm"* (ISO, 2018b). Exposure is defined as *the "state of being in an operational situation that can be hazardous if coincident with the failure"* (ISO, 2018b). Severity is defined as an *"estimate of the extent of harm to one or more individuals that can occur in a potentially hazardous situation"* (ISO, 2018b).

In order to identify the parameters for severity, exposure and controllability ratings, literature review, review of existing standards and accident data were conducted. The rule-set was further re-calibrated (like introduction of new parameters, removal of certain parameters etc.) based on the feedback from international functional safety experts.

Additionally, one of the key parameters identified for the controllability rating was vehicle speed. To establish the rule-set for controllability (with respect to speed of the vehicle), the author performed a driving simulator study involving 44 participants (**study two**) where participants experienced three different speeds (low, medium and high). The details of the study are discussed in appendix three. The objective method created was then applied on to a real-world example (low-speed automated driving system, e.g. pod) via a series of workshops.

4.1.3.2. **RO 6:** *To determine the ability of the developed rule-set for HARA in improving the reliability of the automotive HARA* **(via workshop studies: study four and study five)**

After the creation of the rules for objective HARA (meeting RO 5), the rule-set was then applied on to a real-world example (low-speed automated driving system, e.g. pod) via a series of workshops to establish its effectiveness. **Study four** and **study five** were conducted to establish the basis of the hypothesis and then validate the rule-set created by the author. Both the studies were conducted with international safety analysis experts (from UK, US, Sweden, Hungary, Australia, Italy, Japan and Germany) in order to ensure that the opinions of the automotive industry experts from different countries were captured.

In order to evaluate the reliability challenges, it was important to gather responses from experts from different backgrounds (country, automotive supply chain hierarchy etc.). In order to tackle this challenge, international conferences were identified as a platform where functional safety experts would come together at the same time. In order to answer the research question 3 (improving reliability of automotive HARA) and to meet the research objective 6, the study required safety experts to perform HARA multiple times and in different groups. Therefore, it was decided to use a "workshop" format with multiple rounds as a way of getting experts to perform the HARA. The workshop was inspired from the World Café method (Fouche and Light, 2011). The world café method was used as it keeps the participants engaged in the task given to them and also allows to mix participants to form new groups (which was essential to evaluate the reliability challenges of the automotive HARA).

The workshops were conducted at international conferences in US and Germany. The workshops in UK and Sweden were conducted by inviting industrial contacts of the author which were made during his time in the industry, and while attending various international conferences. Study four was an initial study to evaluate the hypothesis of the existing reliability challenges involved with automotive HARA process. Study five consisted of a series of workshops conducted in Sweden, Germany and the UK. The countries hosting the workshops (and conferences) were selected on the basis of their active work in the area of functional safety and automated driving. Feedback from safety experts was captured from each of the workshops and was incorporated in the rule-set and the experiment design of the subsequent workshops. Study four and study five are discussed in greater detail in chapter 8.

4.2. Driving simulator studies

A driving simulator was used as a tool for studies (discussed in this thesis) to meet the research objectives in both the themes of this thesis (trust and testing), general information about the driving simulator is discussed in this section. More detailed information (specific to each study) is discussed in chapter 5 (for trust studies: study one and study six) and appendix 3 (for controllability study: study two).

Driving simulators offer benefits over real-world driving studies. Specifically, relevant for this thesis, the principal reasons for choosing driving simulator as a tool for this research were:

- Putting the participants in a driver-in-the-loop environment where they are immersed in the simulation environment leading them to behave in a similar manner as real world driving (Underwood et al., 2011)
- Possibility of putting the participants in a dangerous / hazardous driving situation while ensuring they do not experience any physical harm

For the driving simulator studies presented in this thesis, WMG's 3xD driving simulator was used (WMG, 2017). The 3xD simulator offers a unique platform for system validation, as well as for driver behaviour studies around autonomous vehicles with varied levels of automation. The WMG's 3xD simulator provides a 360° Field of View (FoV) which helps immerse the participants in a driver-in-the-loop simulation environment. Apart from the reasons for choosing a driving simulator (mentioned above), the choice of using the 3xD simulator was the ability to control the simulation (i.e., simulation entities like ego vehicle, other vehicles, environment, pedestrians etc.) in real-time. Details about the WMG 3xD simulator have been discussed in Appendix 2.

4.2.1. Driving Scenarios

Scenario creation in the 3xD simulator is a two-step process:

- World generation: this step creates the base map for the simulation environment which involves the road layout, traffic light layout and buildings (Figure 4.3)
- Adding scenario entities to the world: this step adds the dynamic entities of the scenario like other vehicles, pedestrians, cyclists etc. (Figure 4.4)

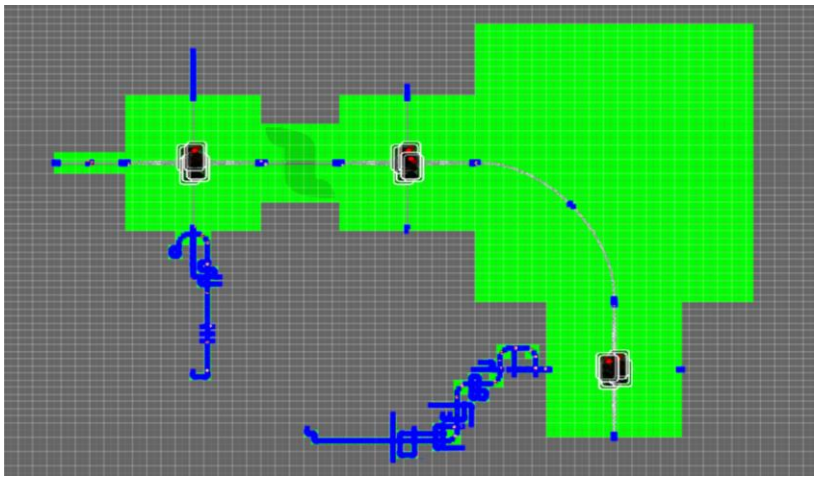

Figure 4.3: World Generator grid environment (with tiles of different road types)

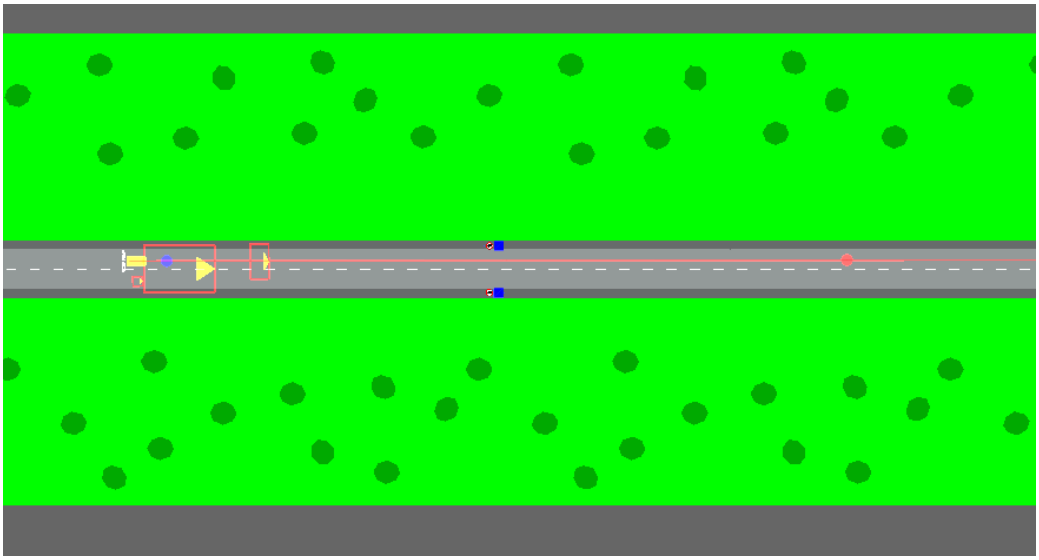

Figure 4.4: Scenario Editor Environment with dynamic elements (vehicle, vehicle path and event triggers)

One of the key learnings on scenario design from study one was the finding that steep turns and roundabouts have the potential to increase chances of participants experiencing simulator sickness. This learning was incorporated in the scenario design for study six in which sweeping turns were used.

4.2.2. Questionnaires used in driving simulator studies

Study one, study two and study six used questionnaires to collect participants' subjective ratings for trust on an automated system, simulator sickness, and the workload during experiment runs, and acceptance of the system under evaluation. Since the trust and

workload questionnaires were used only for the studies evaluating trust (study one and study six), their details have been discussed in chapter five.

Self-reported questionnaires are used to collect subjective data from people in a time and cost efficient manner. However, it is important to ensure that correct language and an easy to understand terminology is used in the questionnaire. A sub-standard questionnaire could potentially confuse the participants by using technical jargon or open ended questions, thus compromising the quality of the collected data.

All driving simulator studies required participants to fill the Simulator Sickness Questionnaire (SSQ) (SSQ) (Kennedy et al., 1993), during and at the end of the study. This was done to evaluate the onset of simulator sickness and monitor participant comfort due to the simulation environment by making a judgement based on the SSQ scores reported between two runs. This was also a way to ensure that the subjective trust and workload ratings weren't affected by the driver experiencing simulator sickness. In case any of the participants felt simulator sickness, the subjective ratings for that run were discarded and not included in the study's analysis.

Simulator Sickness Questionnaire

In order to ensure that ratings collected during the studies correspond to participant's true perception of the system, it is important to ensure that the participants are in a good state of mind. During any driving simulator study, there is always a possibility for the onset of simulator sickness which can affect any subjective ratings provided by participants rendering them unusable for study analysis as they would be influenced by simulator sickness. Therefore, it is important to measure the onset of simulator sickness during any driving simulator study at different points in time. In order to measure simulator sickness, the author used the well-established Simulator Sickness Questionnaire. Participants were asked to fill the SSQ after experiment runs in study one (trust and static knowledge), study two (controllability and speed) and study six (trust and dynamic knowledge). In case there were three or more moderate ratings after any experiment run, the study was stopped.

	Rating			
SSQ Symptom	**None**	**Slight**	**Moderate**	**Severe**
General discomfort				
Fatigue				
Headache				
Eyestrain				
Difficulty focusing				
Increased salivation				
Sweating				
Nausea				
Difficulty concentrating				
Fullness of head				
Blurred vision				
Dizzy (eyes open)				
Dizzy (eyes closed)				
Vertigo				
Stomach awareness				
Burping				

Figure 4.5: Simulator Sickness Questionnaire (adapted from (Kennedy et al., 1993)*)*

4.3. Ethical and practical considerations

4.3.1. Ethical considerations

Since study 1-6 involved human participants, ethical approval was secured from University of Warwick's Biomedical & Scientific Research Ethics Committee (BSREC). All data gathered from the studies was treated in accordance with the University of Warwick's Data Protection Policy and participant's confidentiality was maintained. Informed consent was obtained from all participants.

Learnings from study one and study two (first two driving simulator studies) were incorporated in the design of study six (third and last driving simulator study) in order to reduce the onset of simulator sickness. These learnings influenced scenario design and simulator setup. From study one and study two, it was noticed that roundabouts (with > 90 degree turns) and sharp cornering manoeuvres increased the onset of simulator sickness. Therefore, the scenario design of study six had only sweeping corners and only two roundabouts with 90 degree turns at slow speed (<20 miles per hour). Additionally, in order to increase the immersion of the participants in the driving simulator, additional seat vibration feedback was given to participants which replicated high frequency road feedback which drivers experience in normal road driving conditions. Due to these modifications in study design, study six had a low drop-out rate.

4.3.2. Practical considerations

Since none of the studies provided monetary compensation to the participants for their participation, care was taken to limit the duration of each of the study runs and the interviews.

4.4. Summary

In this chapter, an overview of the various stages of the research undertaken to answer the research questions and meet the research objectives has been presented. While illustrating the two themes of the thesis, the research methods adopted for each of the individual themes have been discussed. Both qualitative and quantitative methods have been used, which together provide a robust approach towards meeting the research objectives of creating and evaluating the effect of informed safety on trust.

CALIBRATING TRUST ON AUTOMATED DRIVING SYSTEMS WITH INFORMED SAFETY[2]

Chapter 5

Having discussed the factors influencing trust on automated driving systems in chapter 3 and introduced the concept of *"informed safety"* in chapter 4, the author will discuss how informed safety aids in developing appropriate trust on automated driving systems. While defining trust in chapter 3, the author mentioned that trust is a history dependent construct, suggesting its dynamic nature. The dynamic nature of trust lends support to the concept of calibration of trust which has been defined as *"the process of adjusting trust to correspond to an objective measure of trustworthiness"* (Muir, 1994). In chapter 3, the author introduced the five stages of calibration of trust: initial phase (stage 1), loss phase (stage 2), distrust phase (stage 3) and recovery phase (stage 4 and stage 5) and the concept of trust calibration number (TCN) as a design goal. The design goal being to minimize trust calibration number by using various intervention strategies to calibrate the trust to a high and appropriate level. There can be various intervention methods to potentially increase /adjust trust in different stages of calibration. One potential intervention method could be the use of knowledge to:

[2] Contents of this chapter have been published in the following publications:

Khastgir, S. et al. (2018a) 'Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles', Transportation Research Part C: Emerging Technologies. Elsevier, 96, pp. 290–303. doi: 10.1016/j.trc.2018.07.001.

Khastgir, S. et al. (2019) 'Effect of Knowledge of Automation Capability on Trust and Workload in an Automated Vehicle: A Driving Simulator Study', in Advances in Human Aspects of Transportation. AHFE 2018. Advances in Intelligent Systems and Computing. Springer, Cham, pp. 410–420. doi: 10.1007/978-3-319-93885-1_37.

a) develop high level of initial trust (in stage 1) by providing accurate static knowledge about the automation capabilities and limitations *(trust in the system)*
b) stem the loss of trust (in stage 2) by complementing static knowledge with dynamic (real-time) knowledge about the automation health, intentions and real-time capabilities *(trust with the system)*

In this chapter, the author discusses the use of knowledge (static and dynamic) as an intervention method to calibrate trust to the "appropriate" level. Thus, meeting the Research Objective (RO) 2 identified in chapter 3 (which answers in part the research question 1 (identified in chapter 3) :

Research Objective 2

"To evaluate the effect of knowledge (static and dynamic) on calibration of trust."

5.1. Introduction

While introduction of automation assumes the removal of human error, in fairness it only shifts the human error from the driver to the designer of the system (Bainbridge, 1983). Muir (1994) has suggested that as the automation capability or reliability increases, trust on automation also increases. Thus, conveying designers' assumptions about the system design (via static or dynamic knowledge) to drivers (users) is essential. However, a mismatch between drivers' perception and expectations about the capability of the automated system, and the designers' assumptions can lead to misuse (due to mistrust), disuse (due to distrust) or abuse of the automated system (Parasuraman and Riley, 1997). Misuse is a situation when the driver uses the automated systems for tasks it was not designed to perform and is caused due to mistrust, thus making the situation more unsafe than manual driving. Disuse is a situation when the driver doesn't use the system in situations where the automation is suitable to use due to distrust or dislike in the system, thus not benefiting from the system. Thus, in order to ensure appropriate use of the system, it is essential to calibrate drivers' trust on the system to the appropriate level. Trust on automation is one of the most important factors influencing use of automation (Lee and Moray, 1992; Muir, 1987; Muir and Moray, 1996; Parasuraman and Miller, 2004; Parasuraman and Riley, 1997; Rudin-Brown and Parker, 2004; Walker et al., 2016).

5.1.1. Knowledge: a factor influencing trust

In order to have appropriate trust, is it important to convey the designer's assumptions about the safe boundaries of the system to the driver. The knowledge of these boundaries provides the ability to have a safe cooperation with the automated system (Beller et al., 2013). In the absence of such knowledge, drivers may not be able to calibrate their trust to an appropriate level (Chavaillaz et al., 2016b; Lee and See, 2004). In chapter three, the concept of "good failures" has been discussed which have negligible impact on trust. As discussed earlier in chapter three, good failures are those whose occurrence is predictable, which allows the driver to be prepared to accommodate for it. Predictability of failures of an automated system comes with knowledge about the true capabilities and limitations of the system.

For complex systems requiring supervision, it has been argued that there is a need for an abstraction hierarchical representation of knowledge of the functional properties of the system (Rasmussen, 1985). The abstraction hierarchy can potentially be done on two fronts. The first category is a whole/part of the system hierarchy, in which the system is viewed as a number of interacting sub-systems working together at different physical levels (Rasmussen, 1985). The second category suggested in Rasmussen's hierarchical knowledge representation is the abstraction of the functionality (Rasmussen, 1985). The physical form of the system represents the lowest level of abstraction. Moving up through the levels, physical functions represents the next level, next is generalized functions, abstract functions forms the penultimate level with functional purpose forming the highest level of knowledge abstraction. The higher abstraction levels do not just represent the abstraction of physical form, they provide knowledge about the control laws for the interactions of the functions at the lower levels. Moving up the abstraction levels provides a purpose of the task for the level below, while moving down the levels provides information about how the task will be achieved.

When put in a driving context, the lower levels of abstraction represent the operational (as per Michon (1985)) driving task (means to a desired end goal) while the higher levels of abstraction represent the tactical and strategic driving tasks (defining the desired end goal). As priority is always given to higher levels of abstraction, a driver has to make a trade-off between the end goal (tactical / strategic goals) and means to achieve it (operational goals), to ensure the means to achieve the goal (lower levels of abstraction) lie within the safe boundaries of the system. In a manual driving task, such a trade-off has clear boundaries and represents a causal system (Rasmussen, 1985). The introduction of automation makes the driving task and the system more complex with blurred boundaries and no simple relationship between function and physical processes making it difficult to represent them as

a causal system. Such systems are referred to as intentional systems. For intentional systems (ADAS and AD systems), decision making requires knowledge about the system, its limitations and the actual input to the system (from the environment) and a top-down approach to control the system in a safe manner (Rasmussen, 1985).

5.1.1.1. Types of knowledge

In chapter three, three types of knowledge about capabilities and limitations of the systems has been proposed based on literature (Bennett, 2017; Biassoni et al., 2016; Feldhütter et al., 2016; Miller et al., 2016; Rasmussen, 1985; Seppelt and Lee, 2007; Xu et al., 2014). These are:

- Static knowledge: Understanding of the functionality of the automated system (intentions behind the design of the system and functionality) (Eichelberger and McCartt, 2014; Larsson, 2012). Static knowledge is administered prior to the driving task and is akin to an owner's instruction manual, however with information at a higher abstraction level. Over time, a person can also build up static knowledge based on experiences. Knowledge of the limitations of the system leads to "good failures" or predictable failures enabling the driver to be prepared for such situations.

- Real time knowledge: or dynamic knowledge about the automated system (e.g. automation health, current state of the automation, near-future intentions of the automation). With the help of real-time information about the automated system health, drivers can be brought back "in-to-the-loop" (Louw and Merat, 2017), as it helps increase their awareness (Banks and Stanton, 2015) and increase transparency in the cooperation between humans and automation (Eriksson and Stanton, 2017b). While in-vehicle information systems (IVISs) are known to have detrimental effect on driving performance (Peng et al., 2014), they have a potential to have a contrasting effect in an automated vehicle as the driver is not actively involved in the driving task. Real time knowledge during repeated driving cycles leads to supplemental static knowledge of the driver about the capability and limitations of the system as it forms part of the consciously imparted knowledge driver brings to the next use of the automated system. Similar to static knowledge, real-time knowledge about the system limitations enables the driver to predict failures (good failures) and be prepared to overcome such situations.

- Internal mental model: Prior beliefs influenced of external sources (e.g. word of mouth, media etc.). Marketing of an automated system can affect the public trust and

perception towards the product. This can potentially backfire if the information provided in marketing material is inaccurate as customers expect the systems to function as advertised (Casner et al., 2016). Inaccurate information can potentially cause over-trust or mistrust in the system. Internal mental model is the pre-conceived notion a person brings to the first use of automation, without any conscious effort to understand the system. While internal mental model is influenced by other sources, static knowledge is consciously imparted to a person prior to the use of automation.

Comparing the presented knowledge classification with Rasmussen's abstraction hierarchies, the author suggests that static knowledge helps adopt a top-down approach, while dynamic knowledge helps adopt a bottom-up approach. Static knowledge further provides the ability to shift the decision making to a higher level or a lower abstraction level depending on the level of dynamic knowledge provided to the driver, i.e. to facilitate the user to more easily transition between levels of the abstraction hierarchy. With the introduction of automation, complexity of system increases, requiring drivers to demonstrate top-down (mean-end) reasoning approach to accommodate for deviations in performance while receiving knowledge about the operational driving parameters (bottom-up knowledge) (Rasmussen, 1985), to demonstrate their knowledge-based behaviour due to unfamiliar nature of the situations (Rasmussen, 1983). The significance of the abstraction hierarchies can be further illustrated by the fact that causes of failures or incorrect function are explained by a bottom-up approach whereas the reasons for the proper function are explained by a top-down approach (Rasmussen, 1985).

Qualitatively, knowledge can potentially be classified into: 1) signals 2) signs and 3) symbols (Rasmussen, 1983). Signals which display time-space sensory data, help the drivers demonstrate skill-based behaviour (based on intuition and experience). While signs indicate towards a stored rule, they do not provide the ability for drivers to process the situation in case a stored rule does not exist in their mental model. Symbols on the other hand represent the relationship between signs and provide the ability for drivers to demonstrate their knowledge-based behaviour and process the information to create a new rule (by shifting the processing to a higher or a lower level of abstraction). Signals, signs and symbols increase situation awareness (SA) of the driver. In chapter three, three levels of situation awareness (SA1, SA2, SA3) have been discussed. Knowledge-based behaviour enables the development of Level 3 SA, enabling the driver to predict future state of the automated system.

5.1.2. Creation of knowledge: identifying failures

While, as described above, providing knowledge to the drivers has a potential of increasing trust, it needs to be stressed that the accuracy of the knowledge provided is key. Inaccurate knowledge plays a detrimental role in development of trust as it takes additional cognitive effort on the part of drivers to re-calibrate their mental model (initially formed in accordance to the inaccurate knowledge) to the true capabilities of the system as they experience the system (Beggiato and Krems, 2013).

In order to create the knowledge of the true capabilities and functionality of the automated system (i.e., to identify failures), it is essential to conduct a thorough verification and validation process. Moreover, due to the safety critical nature of ADAS and AD systems, their deployment needs to be preceded by extensive testing to establish their safety level and performance boundaries (Sepulcre et al., 2013). In chapter three, it was suggested that the identification of failures helps classify them as "good failures" as it provides a level of predictability about them and thus do not have a detrimental effect of trust (Lee and See, 2004). However, knowledge creation about the capabilities and limitations of ADAS and ADS faces reliability challenges (discussed in chapter six) and validation challenges which include challenges in test methods and test setup (discussed in chapter two, six and appendix A2). The approach to knowledge creation reliably which is an important part of the process of development of trust will be discussed in detail in chapter seven and chapter eight.

While defining trust in chapter 3, it was proposed that trust is a history dependent construct, suggesting its dynamic nature. The author adopts the definition of calibration of trust as *"the process of adjusting trust to correspond to an objective measure of trustworthiness"* (Muir, 1994). In chapter three, five stages of calibration of trust were introduced. These are: initial phase (stage 1), loss phase (stage 2), distrust phase (stage 3) and recovery phase (stage 4 and stage 5). There can be various intervention methods to potentially increase/adjust trust in different stages of calibration. In this chapter, the use of static and dynamic knowledge as an intervention method in the process of calibration of trust has been discussed.

5.2. Research Objective

Many authors have studied the effect of reliability (or automation capability) on trust on the system (Chavaillaz et al., 2016b; Muir, 1994; Muir and Moray, 1996), suggesting that with increased reliability, trust increases. However, there is no published research on the effect of static knowledge of automation capability on trust in a driving context (both *"trust in the*

system" and *"trust with the system"*). With the help of driving simulator studies, the Research Objective (RO) 2 identified in chapter 4 has been answered in this chapter:

"To evaluate the effect of knowledge (static and dynamic) on calibration of trust"

In order to answer the above mentioned objective, two driving simulator studies were planned and conducted. The first study evaluated the relation between static knowledge and trust (discussed in section 5.3) and the second study evaluated the relation between dynamic knowledge and trust (discussed in section 5.4).

5.3. Calibrating Trust with Static Knowledge (study one)

5.3.1. Method

5.3.1.1. Driving Simulator

The experimental study was conducted in WMG's 3xD simulator for Intelligent Vehicles at the University of Warwick, UK (WMG, 2017). The 3xD simulator consists of a Land Rover Evoque Built-Up Cab (BUC) which is housed inside a cylindrical screen of 8 m diameter and 3 m height. The cylindrical screen provides a 360º field of view for the driver siting inside the BUC. A push button (with a backlight) (akin to an emergency stop button within a highly autonomous vehicle) was connected (hardwired) to a Raspberry Pi 2 board which in turn was connected to the 3xD simulator through a TCP/IP client-server interface. When the participants pressed the button, the backlight switched-off and the vehicle applied emergency braking and came to a stop. When the participant pressed the button again, the emergency brake was released and vehicle continued to drive in autonomous mode, with the backlight glowing again. This setup enabled a true user in the loop simulation platform, with the user being able to transition in and out of autonomous driving mode anytime they desired, rather than only at predefined, scripted simulator events.

5.3.1.2. Participants

Ethical approval for the experiment was secured from the University of Warwick's Biomedical & Scientific Research Ethics Committee (BSREC) (REGO-2015-1746 AM02). Fifty six participants (16 female and 40 male) were recruited for the study via email invitations. The mean age of the participants was 36.29 years (S.D. = 12.82 years). All participants were required to have a valid, UK full driving license and be at least 21 years of age. The average driving experience of the participants was 14.29 years (S.D. = 13.73 years). The participants' assignment was counter balanced among three groups which were:

1) control group 2) low (20%) capability automation 3) high (80%) capability automation. The difference in automation capability is described in section 5.3.1.3.2. In addition to the study group, five additional participants were part of the pilot study group to fine tune the study. This was done as the mentioned study was the first study to be conducted on the WMG 3xD simulator and simulator and scenario design needed to be re-calibrated based on the user feedback. Informed consent was obtained from all participants.

Out of the 56 participants who took part in the study, eight participants were not able to complete the study due to simulator sickness and technical issues while running the driving simulator. The 48 participants who completed the study were assigned to three groups (see Table 5.1).

Table 5.1: Study design: participant groups

	Control Group: Without knowledge		**Group 1:** Low capability automation	**Group 2:** High capability automation
Number of Participants	8	7	21	12
Run 1	Low capability automation	High capability automation	Without knowledge	Without knowledge
Run 2	High capability automation	Low capability automation	With knowledge	With knowledge

5.3.1.3. Study Design

The experiment was designed as a 2 x 2 mixed factorial design with automation capability as the between-subject factor, and knowledge of the automation capability as a within-subject factor. For the control group, automation capability was used as a within-in subject factor to evaluate whether trust increased with experience without providing any knowledge to the driver (participant) about the automation capability. As a part of the study, each participant was driven in automated mode (SAE Level 4 as per SAE J3016 (SAE International, 2018)) twice and witnessed five hazardous incidents during each complete run. Since the study was evaluating SAE Level 4 automation, participants were asked to sit in the front passenger's seat and hold the emergency stop button in their hands. Such an arrangement also ensured that the participants could only use the button (instead of brake pedal) to stop the vehicle. They were further informed that when the emergency stop button was pressed, the vehicle will apply emergency brakes and will need to cover the braking distance depending on the speed of the vehicle. In cases where the participant met with a simulated accident, the run ended abruptly. The driving simulator route for the experiment involved a drive around the University of Warwick campus. Each complete run lasted around 10 minutes. The route around University of Warwick was chosen to provide a better immersive environment for

the participants as most of them were familiar with the university campus. Additionally, the University of Warwick route in the 3xD simulator has photo-realistic imagery and realistic road feedback (vibration) due to a LiDAR scan input which forms the base for the simulation environment. The speed of the automated vehicle was according to the speed limits set on the campus map, ranging from 10-30 miles per hour.

In order to overcome the lack of real-world consequences often experienced by simulation participants, who can easily choose not to react as they might if their own life were in jeopardy (as in real-world), the study had a gamification aspect to it. The game gave participants a goal during the experiment run and added an element of risk to the study (Table 5.2). Both these factors have been discussed in chapter 3 as being essential to evaluate development of trust on automated system. Participants were awarded 1 point for every second they spent in automated mode. Every time they pressed the button, the button press was classified as a "correct stop" or an "incorrect stop". For every correct stop they were awarded a bonus of 200 points and for every incorrect stop, a penalty of 200 points. Before the run, they were further provided information about what defined a correct and an incorrect stop. A correct stop was one where the participant correctly identified that the automated system wouldn't be able to handle the situation, prompting the participant to intervene and press the emergency stop button. An incorrect stop was one in which the participant pressed the emergency stop button and brought the vehicle to standstill, even though the automated system was capable of handling the situation. Additionally, in case any participant crashed (met an accident), a penalty of 10000 points was given and the experiment run came to an end.

An extremely high penalty was added for a crash to add a high degree of risk and motivate participants to avoid crashing the vehicle as perceived risk influences driver's interaction with the automated system (Eriksson et al., 2017). The penalties were added to get the participants to react in a similar manner as if they were in real danger. The participants were asked to maximise their score. However, the score was not a variable within the study. It was more of a mechanism to encourage engagement in the task. Participants were provided information about their score after the study was completed. Participants were given two objectives: 1) avoid crashing the vehicle by pressing the button (emergency stop) 2) maximize time spent in automated mode. They were asked to press the button only if they felt that the automated system couldn't handle the situation or if they felt unsure about the automated system's performance.

Table 5.2: Scoring criteria for study (gamification)

Type of Action	Points
Automated mode	1 / second
Correct Stoppage of the automated vehicle	+200
Incorrect Stoppage of the automated vehicle	-200
Crash	-10000

5.3.1.3.1. Hazards

In order to choose the five hazardous events, a hazard analysis of an automated vehicle was conducted as per the ISO 26262 (ISO, 2011a) functional safety process. Five different automated vehicle functions were identified and a hazard was identified for each of the functions (Table 5.3). For each hazard, a hazardous event was identified which was created in each of the driving scenarios in the experiment runs in the 3xD simulator. The hazard and hazardous event identification was done by independent safety experts. One of the factors influencing the selection of the hazardous events was the ability to create the events in the 3xD simulator.

Table 5.3: Description of five hazardous events

Function	Hazard	Hazardous event description
Braking	Lack of Braking	Pedestrian suddenly changes direction and comes in front of ego vehicle (automated vehicle)
Torque	Excessive torque – excessive acceleration	Vehicle approaching round-about and accelerates instead of braking
Object Detection	Blind-spot and delayed object detection	Another vehicle in perpendicular lane comes in path of the ego vehicle suddenly
Path Planning	Not following rules of road	Ego vehicle joins a roundabout while another vehicle is still in the roundabout and has right of way.
Object Detection	Compromised detection due to environmental factors	In foggy/rainy weather, ego vehicle is not able to detect traffic lights within the specified range.

5.3.1.3.2. Automation Capability

Two levels of automation capability were used in the study: 1) low capability automation 2) high capability automation. The difference between the two systems was based on the ability of the automated system to tackle the five hazardous events mentioned in section 5.3.1.3.1. Low capability automated system was able to handle one out of the five hazardous events,

requiring the driver to intervene in four hazardous events to ensure safe performance of the vehicle. High capability automated system was able to handle four out of the five hazardous events, requiring the driver to intervene in only one hazardous event situation to ensure safe performance.

5.3.1.4. Procedure

When participants arrived for the experiment, they were initially briefed about the experiment following which informed consent was taken from each participant and they were asked to fill in a demographic questionnaire. Before the start of the study runs, each participant was given a trial run (on a route different from the one used for the study runs) on the driving simulator with the author seated next to the participant, to familiarize the participant with the visuals, motion feedback, experience of sitting inside a car within a simulator and using the button to apply emergency brake on the vehicle. Participants were told that they can ask for as many trial runs as they wish, in order to make them comfortable with the simulator environment. Each trial run was of five minutes in length. While most of the participants requested only one trial run, some participants requested for an additional (second) trial run. After the trial runs, participants were asked whether they would like to continue the study. In the case that the participant agreed, each participant experienced two experiment runs of around 10 minutes each. Before the second run (for group 1 and group 2), participants were provided knowledge about the capabilities of the automated system. Commentary was read out to them via a prepared script. Effort was put into the preparation of the script in order to avoid introducing any experiment bias. The script was reviewed by three independent human factors experts.

For the control group, participants were told that in the two runs, they will experience automated control systems from two different suppliers. No other information about system capabilities was given. However, before the second run, it was reiterated that the participants will now experience a different automated control system from a different supplier. Such a design of the control group was implemented to check if there was any changes in the trust levels due to experience. Eight out of 15 participants in the control group experienced low capability automation in their first run and high capability automation in their second run. The remaining seven participants experienced the runs in the reverse order. Add the end of each experiment run, participants were asked to fill a trust rating questionnaire (discussed in section 5.4.1.5.2) and Simulator Sickness Questionnaire (SSQ) (Kennedy et al., 1993).

5.4.1.5.1. Imparting Static Knowledge

Knowledge was imparted to the participants via a prepared script which included illustrations regarding the automated systems' capability and limitations. Special care was taken to ensure that participant's mental model was informed so that they understood the functioning of the system in a lay-man language to ensure higher level system understanding. This was particularly important in order to ensure they were imparted with knowledge-based behaviour, as compared to rule-based or skill-based behaviour. The knowledge imparted would enable them to deal with the unfamiliar situation by transferring the cognitive task to a higher level or a lower level of abstraction in search of an existing rule or intuition of their mental model (Rasmussen, 1985). In the automated driving context, the significance of knowledge-based behaviour is further emphasized as it helps a driver adopt a means-end approach to execute the appropriate human intervention needed for the task. The following scripts were used to impart knowledge to the participants.

Script 1: *"The automated control system from the two suppliers are based on camera based sensors and each automated control system will be trialled in separate runs in the sim. However, due to cost pressures, they have chosen a single low quality camera with reduced field of view.*

Vision based systems are dependent on the quality of the camera used. Due to cost pressures, the supplier has compromised with the accuracy of the camera used for the vehicle. In this vehicle, a lower grade camera has been used. Lower grade cameras are vulnerable to environmental factors and image recognition degrades with lower visibility. E.g., certain cameras find it hard to detect objects in rain, snow or fog or at certain times of the day due to image washout (Figure 5.1). In your drive today, you might have witnessed brigh sunlight or rain. You have the luxury of using sunglasses, wipers etc. However, Camera doesn't have that. It has been found that light colour objects against a bright sky is difficult to detect. This was the case in the recent Tesla Model S crash (NHTSA, 2017a) *where the white rear end of the truck was not detected against the bright sky."*

Figure 5.1: Camera view while driving in fog
(Image source: https://www.flickr.com/photos/kubina/2160242894; date accessed: 2017-12-04)

Script 2: Obstacle detection, e.g., pedestrian detection or vehicle detection is a challenge for on-board sensors as it requires sensor fusion between different sensor readings, like RADAR data and camera data. This is dependent of the quality and amount of data received by the sensor fusion control unit. The supplier is using only camera based systems. No other sensor systems have been used.

Camera based systems are unable to process the images for sudden obstacle detection (e.g. pedestrian jumping on the street) due to high computing power. In optimum lighting, camera based systems are suitable when steady performance is required (e.g. detecting lane

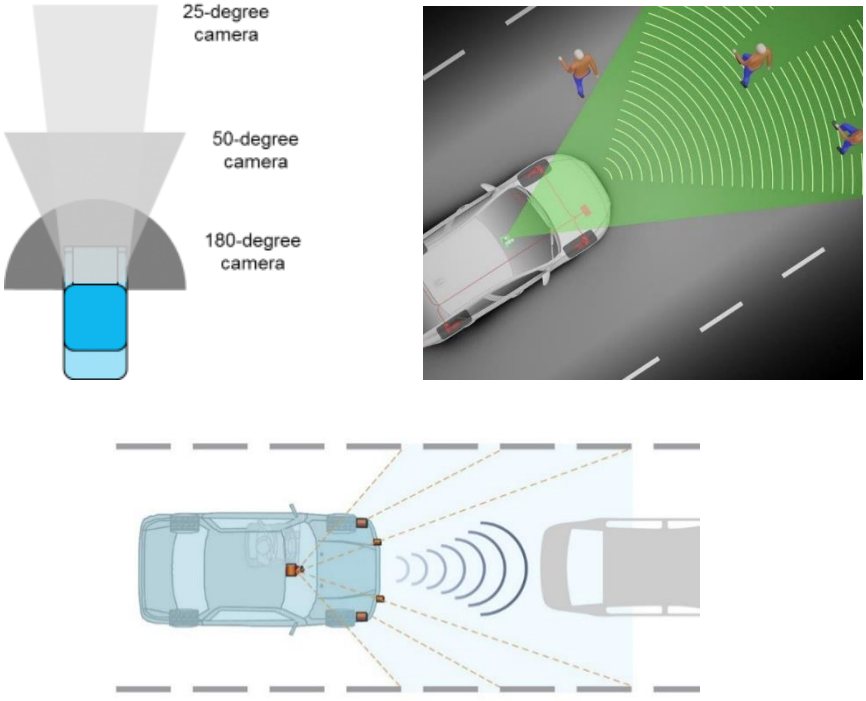

Figure 5.2: Field of view of camera based detection systems

markings, traffic lights or speed signs). As drivers we know that we are supposed to scan all around our vehicle to anticipate sudden obstacles.

In order to avoid sudden obstacles, the vehicle needs either a good camera sensor with wide field of view to sense the oncoming obstacle or multiple cameras to give it a wide field of view (Figure 5.2). Some camera sensors have a wide field of view or multiple cameras can be used to obtain a wide field of view. However, since this supplier is using a single, low quality camera, dynamic object detection may be compromised.

Script 3: "While, automated vehicles have a repeatable and predictable behaviour, their behaviour is "programmed" by human engineer. Every vehicle before being released to market undergoes rigorous testing. However, it is possible that sometimes a programming bug introduced by a human error manifests itself into a larger failure. The rules of the road are pre-programmed into the automated control system. The automated system in your next run is a pre-production control system and is still undergoing testing. While previous test results have been extremely positive, I advise you to take caution. An example of this might be that as a driver, we know that if a pedestrian is standing next to a zebra crossing, they have the right of way (Figure 5.3). However, for a camera system, he/she will only be a pedestrian with unknown intention. In this example the automated control system wouldn't know the rules of the road and will not have the understanding of the priorities.

Another rule of the road that we as drivers are used to is the priorities at roundabouts and junctions (Figure 5.3). Imagine a person is given a driving license when he/she doesn't know the rules of the road. Not only its dangerous for him/her, it is hazardous for the traffic around."

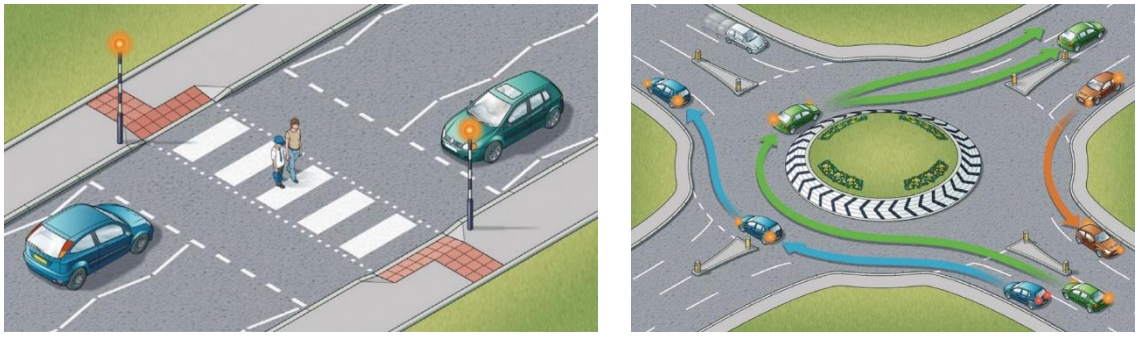

Figure 5.3: Rules of road: rule 19 (left) and rule 185(right). (DfT, 2017)

In the above examples, effort was made to differentiate between knowledge and rule-based behaviours. Simple rules are comparatively easy to convey to participants, for example in Figure 5.1, a rule would be 'automated system will not work in fog'. However, there is no

understanding why it will not work (e.g. *image recognition degrades with lower visibility* which was provided as a part of the script). Knowledge about other similar situation where the camera may not work was also provided via the script (…*hard to detect objects in rain, snow or fog or at certain times of the day);* (*You have the luxury of using sunglasses, wipers etc. However, Camera doesn't have that. It has been found that light colour objects against a bright sky is difficult to detect. This was the case in the recent Tesla Model S crash where the white rear end of the truck was not detected against the bright sky*). By trying to impart knowledge the participant can envisage their own varied and numerous situations where the automated system might act unexpectedly.

5.4.1.5.2. Trust Questionnaire

Trust is a subjective construct and has been measured via different rating scales in literature (Jian et al., 2000; Muir and Moray, 1996). As discussed in chapter 3, the author classified trust into two aspects: *"trust in the system"* and *"trust with the system"*. A subjective rating scale was used and participants were asked to draw a line across a 100 mm box to indicate their level of trust (c.f. (Muir and Moray, 1996; Rajaonah et al., 2006)). Before being asked to rate different trust levels, participants were briefed about the difference in the different types of trust via a prepared script which included examples (was read to the participants as well as given in text form) to highlight the difference between *"trust in the system"* and *"trust with the system*. Existing rating scales like Jian's scale (Jian et al., 2000), couldn't be used as they don't classify trust into the two components discussed in chapter 3 (as they treat trust as a single construct). In order to explain the two different concepts of trust, participants were briefed using an example of a mobile phone and call service provider. The following text was used for the explanation:

"Trust in the system means that you have trust on the capabilities of the system and in its ability to do what it is supposed to do as advertised to you. In other words, it does what it says on the box. Trust with the system means that you are aware of the limitations of the systems and you adapt your use of the system to accommodate for the limitations in order to get maximum benefit from the system.

For example, if you buy a mobile phone, you have trust ***in*** *the systems about its advertised capabilities. You develop trust* ***with*** *the system once you start using it and understand its limitations. Ability to work with limitations guides your trust* ***with*** *the system. For the mobile phone and the call service provider you have, you get call drop-outs in certain part of our house and not in another part of your house. You would adapt your usage of the mobile phone by making calls only when you are in a part of the house where you know call*

connection service is good. This is an example of you acknowledging the limitations of the system, adapting your usage and developing trust with the system"

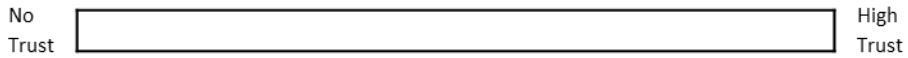

Figure 5.4: Subjective (Trust) rating scale (100 mm) (c.f. (Muir and Moray 1996; Rajaonah, Anceaux, and Vienne 2006)

On the trust scale (Figure 5.4), a 0% rating suggested very low trust and 100% suggested very high trust. As trust is a continuum, any value in between 0 -100 suggests that the participant had partial trust.

5.3.2. Results

5.3.2.1. Trust levels

The average *"trust in the system"* for low capability automation increased substantially from 32.4% to 65.4 %, with the introduction of knowledge about the system capabilities and limitations (Figure 5.5). While an increase in *"trust in the system"* rating with the introduction of knowledge was seen for high capability automation from 54.2% to 70.5% also, the effect was comparatively lower. It is interesting to note that with the introduction of knowledge about the automated system's capabilities and limitations, both median and mean values for *"trust in the system"* for low-capability and high-capability automated system were similar (Figure 5.5). In the low capability automation group, barring two participants out of the 21 participants, all participants showed an increase in trust in the system with the introduction of knowledge (Figure 5.6). High capability automation group also showed a similar trend. The box-plots for trust in the system illustrate a higher convergence in trust ratings with the introduction of knowledge, potentially due to appropriate calibration of trust level (Figure 5.5).

A repeated measures ANOVA was conducted for the *"trust in the system"* and *"trust with the system"* ratings with automation capability as the between factor variable and knowledge as the within factor variable. The introduction of knowledge about the automation capabilities and limitations had a highly significant statistical effect on the level of *"trust in the system"*, $F(1, 31) = 33.712$, $p = 0.000002$ with a $\eta_p^2 = 0.521$, suggesting 52.1% of the variance being associated with the introduction of knowledge. While the introduction of knowledge didn't have an interaction effect with automation capability, $F(1, 31) = 3.846$, $p = 0.059$ ($\eta_p^2 = 0.11$). Therefore, there was no effect of automation capability on trust in the system ratings when knowledge was introduced.

While the average *"trust with the system*" changed with the introduction of knowledge (Figure 5.7), the effect was statistically insignificant, $F(1, 31) = 3.652$, $p = 0.065$ with a $\eta_p^2 = 0.105$. There was no interaction effect between knowledge and automation capability for trust with the system ratings, $F(1, 31) = 0.742$, $p = 0.396$ ($\eta_p^2 = 0.023$).

In order to negate the effect of experience on trust ratings, a repeated measures ANOVA was performed on the control group. The effect of the runs was statistically highly insignificant on the level of *"trust in the system"*, $F(1, 13) = 0.105$, $p = 0.751$ with a $\eta_p^2 = 0.008$. There were no interaction effects between the runs and the two control groups, $F(1, 13) = 0.020$, $p = 0.89$ ($\eta_p^2 = 0.002$).

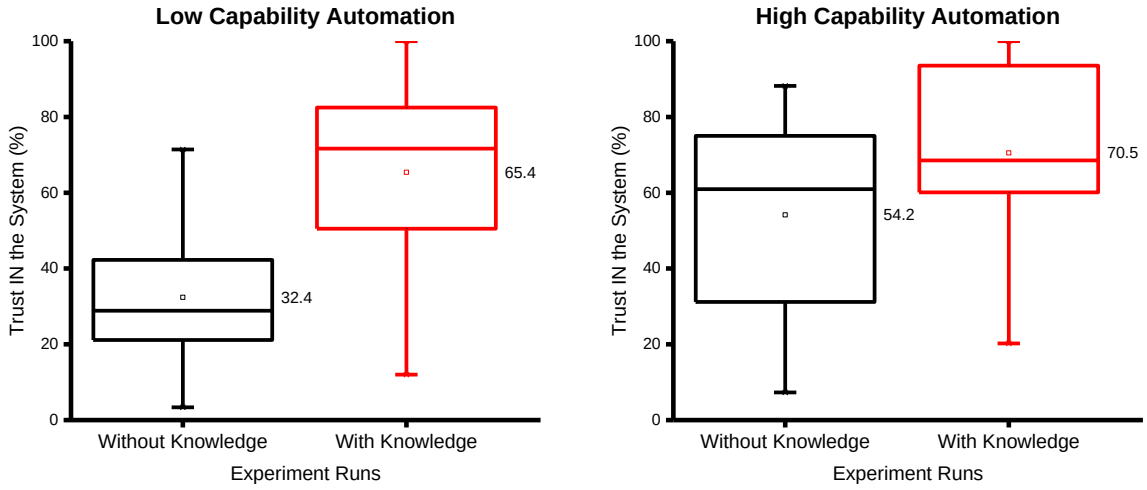

Figure 5.5: Box-plots of Trust-In the system ratings (highlighting average trust ratings) (central dot represents average value)

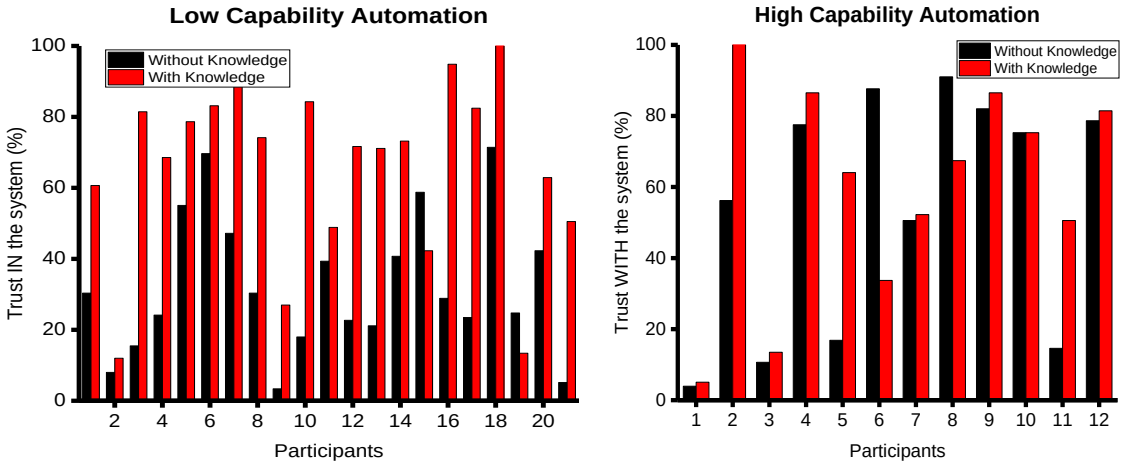

Figure 5.6: "Trust in the System" level of individual participants for low capability and high capability automation

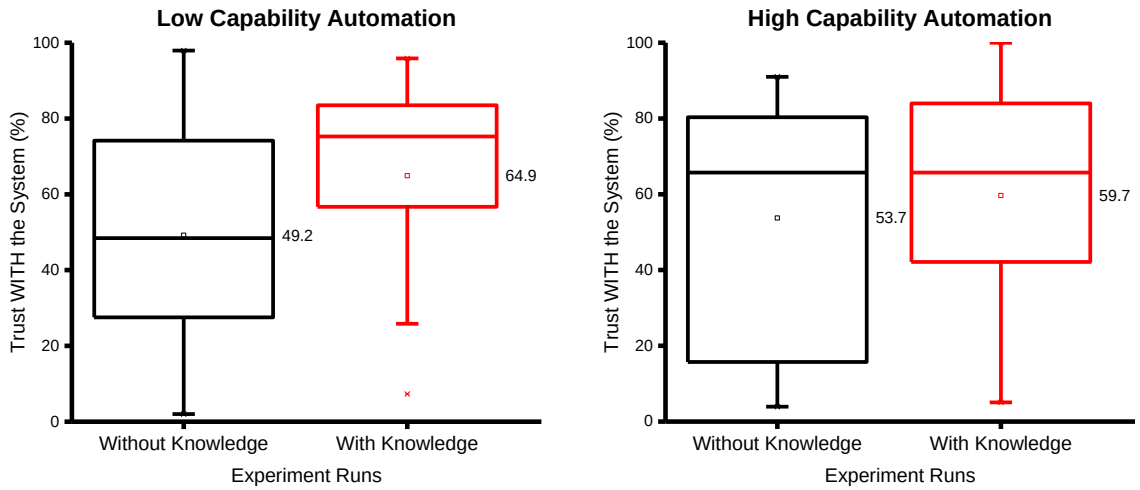

Figure 5.7: Box-plots of Trust-With the system ratings (highlighting average trust ratings) (central dot represents average value)

5.3.2.2. False presses

While the introduction of knowledge about system capabilities and limitations increased trust in the system for both low and high capability automation, it had contrasting effect in the two groups in terms of number of false presses. In the context of this study, a false press is defined as a button press in a situation which could be handled by the automated system, indicating distrust in the system.

For low capability automation, the average number of false presses increased significantly from 0.47 to 2.67 with the introduction of knowledge. On the contrary, for high capability automation the average number of false presses decreased from 1.73 to 1.36 with the introduction of knowledge (Figure 5.8). The outlier data from the box-plot were removed for mean calculation. This meant one data point each from the two runs for high capability automation was removed. There were no outliers in the data set for low capability automation group.

As discussed in section 5.4.1.5.1, for the low capability automation group, participants were given lot of knowledge due to limited capability. One of the potential reasons for the contrasting results between the two groups could be the amount of knowledge provided in the low capability automation group and the participants' ability to process all the knowledge, develop accurate mental model and display knowledge-based behaviour. However, higher trust ratings with introduction of knowledge suggest that knowledge-based behaviour was displayed. Another potential reason for the contradictory results could be the

lack of dynamic (real-time) knowledge provided to the participants (discussed in section 5.3.3).

A paired-sample t-Test was conducted to assess the significance in the number of false presses with the introduction of knowledge. For low capability automation, there was a statistically significant difference in the number of False Presses (FP) for without knowledge run (M = 0.47, SD = 0.60) and knowledge run (M = 2.67, SD = 1.65); t (20) = -6.398, p = 0.000003. For high capability automation, the number of False Presses (FP) for without knowledge run (M = 2.41, SD = 2.79) and knowledge run (M = 1.67, SD = 1.43) was statistically insignificant; t (11) = 0.792, p = 0.445.

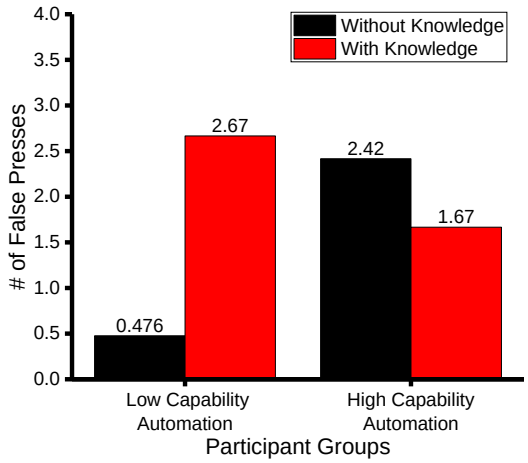

Figure 5.8: Average number of false presses

5.3.2.3. Accidents

In the context of this study, an accident is defined as a collision of the ego vehicle (automated vehicle) with other entities (vehicles, pedestrians or cyclists) in the scenario or if the own vehicle doesn't follow the traffic light rules. Introduction of knowledge about the automated system capability had similar effect on the average number of accidents for both the automation groups. For low capability automation, the average number of accidents reduced significantly from 1 to 0.38 with the introduction of knowledge (Figure 5.9). For high capability automation, the average number of accidents reduced slightly from 0.58 to 0.42 (Figure 5.9). It is interesting to note that most of the accidents were caused to due to late interventions rather than absence of interventions. This may be explained due to lack of accurate situation awareness about scenario handling capabilities of the automated system during the automated driving scenario which could potentially be due to the lack of dynamic

knowledge of the participants. A paired sample t-Test was conducted to assess the statistical significance in the number of accidents with the introduction of knowledge. There was a statistically significant difference in the number of accidents between the without knowledge (M = 1, SD = 0) and with knowledge (M = 0.38, SD = 0.49) conditions; t (20) = 5.701, p = 0.000014, for low capability system.

Similar to the false presses, the number of accidents for without knowledge (M = 0.5, SD = .52) and with knowledge runs (M = 0.42, SD = 0.51) conditions for high capability automation was insignificant; t (11) = 0.321, p = 0.754.

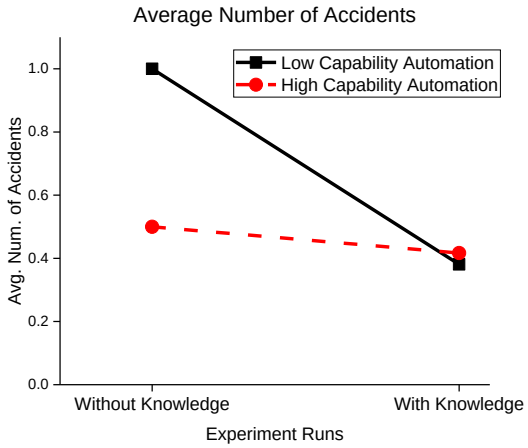

Figure 5.9: Average number of accidents

5.3.2.4. Workload

While conducting study one, the author received qualitative feedback from the participants suggesting that there was an increase in their workload during the driving task when static knowledge was provided to them. To evaluate this suggestion scientifically, workload experienced by participants was also measured during the latter part of the study one using the Overall Workload Scale (OWS) (Hill et al., 1992). OWS scale requires participants to rate their workload on a scale of 0-100 with intervals of five. In the study presented, participants were not exposed to any secondary task. Therefore, the measure of the internal workload (measured using OWS) can be considered as the total workload of the driver. The extended study had 19 participants assigned into three groups: 1) control group (N = 4) 2) low capability automation group (N = 6) and 3) high capability automation group (N = 9). Table 5.4 depicts the average workload levels for the two automation capability groups.

Table 5.4: Average Workload levels

Group	Average Workload level (out of 100)	
	Without knowledge	With knowledge
Low capability automation	24.2	61.7
High capability automation	55.5	61.1

Figure 5.10 depicts the workload ratings given by the participants in different groups. Interestingly, when knowledge about the automation capability was provided, the increase in workload in the low automation capability group was larger than in the high automation capability group. This result may be due to the increased number of interventions required to ensure safety while using a low capability automated system and more effort required to recalibrate the driver's mental model of the system capability based on the knowledge provided.

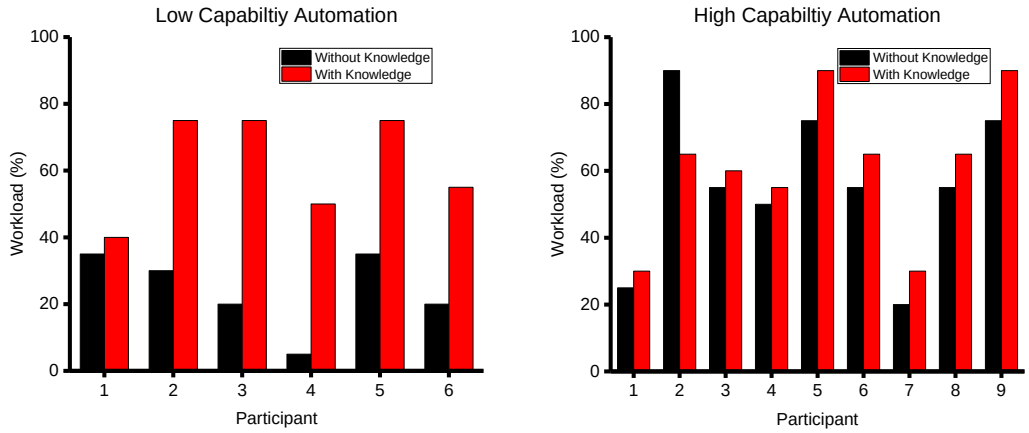

Figure 5.10: Workload ratings

A repeated measures ANOVA was conducted for the workload ratings with automation capability as the between factor variable and knowledge as the within factor variable. The introduction of knowledge about the automation capabilities and limitations had a highly significant statistical effect on the workload level, $F(1, 13) = 32.619$, $p = 0.000072$ with a $\eta_p^2 = 0.715$, suggesting 71.5% of the variance being associated with the introduction of knowledge. However, the introduction of knowledge had a statistically significant interaction effect with automation capability, $F(1, 13) = 17.956$, $p = 0.001$ ($\eta_p^2 = 0.580$). For post-hoc analysis, paired sample t-Tests were conducted for low capability and high capability automation. There was a significant difference in the workload levels for low capability automation between the runs without knowledge ($M = 24.17$, $SD = 11.58$) and the runs with knowledge ($M = 61.67$, $SD = 15.38$); $t(5) = -5.326$, $p = 0.003$. There was an insignificant difference in the workload levels for high capability automation between the

runs without knowledge (M = 55.55, SD = 22.83) and the runs with knowledge (M = 61.11, SD = 21.47) conditions; t (8) = -1.37, p = 0.206. This suggests that introduction of static knowledge increases workload levels in the case of low capability automation, while there is no significant effect in the case of high capability automation. Thus providing us another reason to ensure the introduction of high capability automated systems along with the identification of their limitations and capabilities which need to be communicated to the drivers.

5.3.3. Discussion

Study one has demonstrated that with the introduction of static knowledge about the system capabilities and limitations, *"trust in the system"* increases, to similar trust ratings for low-capability and high-capability systems. These results differ from the study in (Helldin et al., 2013) and (Hergeth et al., 2017). While these studies did provide some feedback about the system boundaries to the drivers, they were unable to instil knowledge-based behaviour as they didn't mention how the system works due to which the driver's higher level mental model could not be made.

It is worth noting that the effect of knowledge on *"trust in the system"* had a statistically highly significant relationship (p = 0.000002), the effect of knowledge on *"trust with the system"* was statistically not significant (p = 0.065). This can be explained by analysing the nature of knowledge provided to the participants. As mentioned in section 5.1.1, knowledge can be qualitatively classified into three categories. In the study presented, participants were provided with only static knowledge about the capabilities and limitations of the systems. While this allowed them to demonstrate their knowledge-based behaviour and helped them calibrate their trust in the system, the lack of system feedback on the real-time state and intention of the system, led to lower levels of trust with the system. This inference is further corroborated by the qualitative feedback from participants who were asked to explain their rating of trust in their own words. One of the participants (participant #20) commented: *"warnings from the car missing"* while other (participant # 40) commented *"no warnings & notification"*. Another participant (participant #37) mentioned: *"I was able to accommodate for the system but it was discomforting… near misses and close calls"*. These comments can possibly explain the high level of workload experience by the drivers with the introduction of static knowledge (in low capability automation group) as dynamic (real-time) knowledge was absent in study one.

In other words, the introduction of static knowledge provided participants the capability to demonstrate top-down understanding as per the abstraction hierarchy levels. However, with the absence of dynamic knowledge, they were unable to get feedback (signs and signals) on

the causes of the failure, subsequently their reasoning capability was limited. Thus, in order to be able to work with the system, i.e. accommodate for the limitations of the system and display their knowledge-based behaviour appropriately, participants also require real-time knowledge (e.g. signals and signs) to move the decision task to a higher or a lower abstraction level in search of pre-existing rules or intuition, similar to a co-pilot in the aviation domain (Eriksson and Stanton, 2017b). Thus, the author suggests that *"trust with the system"* is potentially influenced to a larger extent by dynamic (real-time) knowledge about the system capabilities and limitation and this subject has been explored in section 5.4.

The introduction of knowledge didn't have an interaction effect with automation capability on trust ratings ($p = 0.059$ for *"trust in the system"* and $p = 0.065$ for *"trust with the system"* ratings). Thus suggesting that similar levels of trust can be achieved if knowledge about the true capabilities and limitations of the systems is provided to the driver.

While due to the study design the control group's trust ratings can't be compared with the low-capability automation or high-capability automation group's trust ratings, they do provide more confidence in the results obtained in the two latter groups. The role of the control group was to either support or negate the hypothesis that any change in trust ratings could be a result of experience. Results showed that automation capability has no interaction effect on experience of the system ($p = 0.89$), thus negating the hypothesis.

5.3.4. Study one limitations

Knowledge has two main facets (which can be imparted through system design): 1) static knowledge (e.g. initial briefing and driving manual) and 2) dynamic knowledge (via as human-machine interface display). In study one (discussed in section 5.3), the author only provided static knowledge to the participants. Study six (discussed in section 5.4) was planned where participants were provided both dynamic knowledge and static knowledge to compare the effect of static and dynamic knowledge on trust ratings.

5.4. Calibrating Trust with Dynamic Knowledge (study six)

In section 5.1.1.1, knowledge has been classified into static knowledge, dynamic knowledge and internal mental model. It is evident from study one (trust and static knowledge), that the introduction of static knowledge while increasing trust levels, also increases workload. In this section, the effect of dynamic knowledge on trust is investigated by means of a driving simulator study. Dynamic knowledge provides real-time information about the automation health and near-future intentions of the system. The main difference with static knowledge is that static knowledge is administered prior to use of the system while dynamic knowledge is administered continuously in real time. Real time status of the system or warnings have a potential to increase safe use by drivers' situation awareness and enable them to display knowledge-based behaviour (Baldwin and Lewis, 2014; Robinson, 1986). However, it is essential to design the dynamic knowledge in a way that is understandable to drivers by ensuring concise and brevity of the knowledge and also be accurate (Riley, 2014; Wogalter et al., 1991). The accuracy aspect of dynamic knowledge is associated with the creation of the knowledge and will be discussed in detail in chapter six and seven which focus on test scenarios. The hypothesis for this study is - *"dynamic knowledge will enable the drivers to work with the system"*. This is based on the feedback from the participants of study one (trust and static knowledge) who mentioned about the lack of warnings.

5.4.1. Method

5.4.1.1. Driving Simulator

Similar to study one (section 5.3), the experimental study was conducted in WMG's 3xD simulator for Intelligent Vehicles at the University of Warwick, UK (WMG, 2017) with a push button (with a backlight) which was connected (hardwired) to a Raspberry Pi 3 board which in turn was connected to the 3xD simulator through a TCP/IP client-server interface. In this study a Raspberry Pi 3 was used instead of Raspberry Pi 2 as the experiment setup required connecting two Raspberry Pis via Bluetooth connection. Bluetooth adapter is available only on Raspberry Pi 3. Participants could engage or disengage automated mode by pressing the button. The button backlight glowed in automated mode and was turned off in manual mode. There were other mechanisms to disengage automated driving mode (i.e. to bring the car in manual mode) which have been discussed in section 5.4.1.4.

5.4.1.2. Human-Machine Interface (HMI) display setup

In order to provide real-time information to the participants, an HMI display was designed and mounted over the central console of the Land Rover Evoque (BUC). A Raspberry Pi screen attached with a Raspberry Pi was used as the HMI display. In order to create the HMI design, GL Studio software was used. The Raspberry Pi and the screen was connected to a laptop on which GL Studio HMI design was designed. The HMI design is discussed in section 5.4.1.5.1 (titled Imparting Dynamic Knowledge). An "http" server was implemented in GL studio and the corresponding executable was executed on the Raspberry Pi attached to the screen (in the BUC). The laptop and the Raspberry Pi screen communicated via the http server-client interface. A second Raspberry Pi communicated with the 3xD simulator via a TCP/IP client-server interface (similar to study one: section 5.3.1.1). The laptop (with http client) could have been directly connected via UDP (User Datagram Protocol which is another type of communication protocol between two processors) to this (second) Raspberry Pi. However, a third Raspberry Pi was required in order to reduce the network load on the Ethernet port of the second Pi. If both UDP (to laptop) and TCP/IP (to simulator) were done from the same Ethernet port (of the second Pi), many communication (message) packets were missed. Therefore, the Ethernet port of second Pi was exclusively used for TCP/IP messages and its Bluetooth adaptor was used to transfer data to the third Pi whose Ethernet port was exclusively used for UDP communication with the laptop. Figure 5.11 shows the schematic of the Raspberry Pi, laptop and the simulator setup.

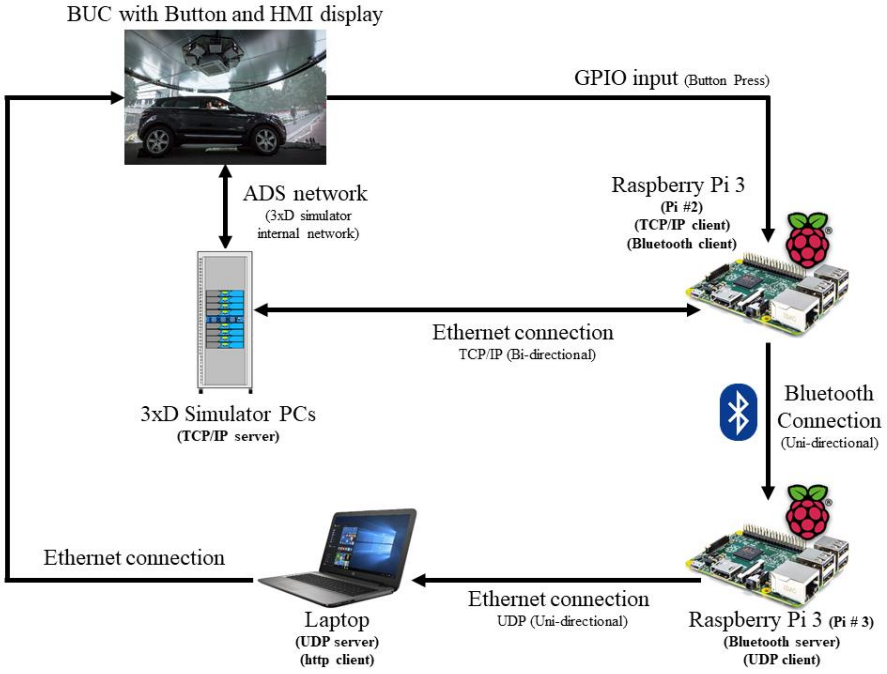

Figure 5.11: Study six setup schematic

5.4.1.3. Participants

Ethical approval for the experiment was secured from the University of Warwick's Biomedical & Scientific Research Ethics Committee (BSREC) (REGO-2015-1746 AM02). Thirty seven participants (six female and 31 male) were recruited for the study via email invitations. The mean age of the participants was 34.81 years (S.D. = 10.39 years). All participants were required to have a valid driving license and be at least 21 years of age. The average driving experience of the participants was 14.69 years (S.D. = 11.78 years). Informed consent was obtained from all participants.

Out of the 37 participants who took part in the study, two participants were not able to complete the study due to the onset of simulator sickness while experiencing the driving simulator. Thirty five participants who completed the study were assigned to four groups (Table 5.5).

Table 5.5: Study 6 design: participant groups

	Group 1: Low capability automation			**Group 2: High capability automation**
Number of Participants	7 (Group L1a)	6 (Group L1b)	8 (Group L2)	14 (Group H)
Run 1	No Knowledge	No Knowledge	No Knowledge	No Knowledge
Run 2	Dynamic Knowledge	Dynamic Knowledge + Static Knowledge	Dynamic Knowledge	Dynamic Knowledge
Run 3	No Knowledge	No Knowledge	-	-

5.4.1.4. Study Design

The experiment was designed as a 2 x 2 mixed factorial design with automation capability as the between-subject factor and knowledge of the automation capability as a within-subject factor. Out of the 35 participants, 13 participants underwent three different runs while remaining 22 participants experienced two runs. With the benefit of hindsight, for future studies, it was concluded to use statistical power analysis to determine the sample size and group size. In order to negate the possible argument that trust ratings would increase due to learning experience, the study design required 13 participants to experience three runs (Group L1a and L1b). Additionally, three study runs required around 90 -110 minutes commitment from the participants. As the study wasn't providing any monetary incentives for participants, author believed that over 90 minutes commitment would be a lot from all 35 participants. Therefore, due to time constraints, 13 participants were asked to undergo three runs, as it was the minimum requirement of the study design. Furthermore, to explore the difference in the effect of only dynamic knowledge and dynamic knowledge along with

static knowledge, the study design differentiated low-capability automation groups into group L1a and group L1b, with the former receiving only dynamic knowledge in run 2 and the latter receiving both dynamic and static knowledge in run 2. No such differentiation was implemented for Group H (high capability automation) with all participants receiving only dynamic knowledge in run 2. As a part of the study, each participant was driven in automated mode twice (group L2 and group H) and thrice (for group L1a and group L1b) and witnessed five hazardous events during each complete run. Participants could disengage automation (i.e. starting driving in manual mode) by either pressing the brake or accelerator pedal, or turning the steering wheel or by pressing a button present inside the vehicle. Once in manual mode, participants could re-engage automation by pressing the button. The button had a backlight which glowed in automated mode indicating mode of the vehicle.

In contrast with study one, the driving simulator route for the experiment involved a drive in a built-up world with sweeping bends to reduce the onset of simulator sickness. The built-up world had urban, rural and motorway sections. Each complete run lasted around 15 minutes. The speed of the automated vehicle was according to the speed limits defined in the built-up world (urban: 30 mph, rural: 40 mph and motorway: 60 mph).

Similar to study one (section 5.3), this study also had a gamification aspect to it, in order to give participants a goal during the experiment run and add an element of risk to the study (Table 5.6). The game design was to overcome the lack of real-world consequences often experienced by participants in a simulator environment, who may not react like they would in real-world if their own life was in jeopardy. In addition, the scoring gave participants a goal for their task (along with ensuring safe operation of the vehicle) and the penalties added an element of risk during the task. Participants were awarded 1 point for every second they spent in manual mode and 10 points for every second they spent in automated mode. Every time participants made a transition from automation to manual or vice-versa, the transition was classified as a "correct" or an "incorrect" transition. For every correct transition they were awarded a bonus of 100 points and for every incorrect transition, a penalty of 200 points. In case they met with an accident or had a traffic rule violation (going through a red traffic light or speeding), a penalty of 4500 points was incurred. Similar to study one, an extremely high penalty was added for an accident and traffic rule violation to add a high degree of risk and to motivate participants to avoid them. Similar to study one, participants were asked to maximise their score. Participants were given two objectives: 1) avoid any accident or traffic rule violation by the vehicle by taking control of the vehicle 2) maximize time spent in automated mode. They were asked to transition to manual mode only if they felt that the automated system could not handle the situation or if they were not confident about its abilities.

Table 5.6: Scoring criteria for study 6 (gamification)

Type of Action	Points
Manual mode	1 / second
Automated mode	10 / second
Correct transition (manual to automated or vice versa)	+100
Incorrect transition (manual to automated or vice versa)	-200
Accident (crash) / traffic rule violation	-4500

5.4.1.4.1. Hazards

In order to choose the five hazardous events, a hazard analysis of an automated vehicle was conducted using the STPA process (discussed in chapter 7). Like in study one (section 5.3), five aspects of an automated vehicle functions were identified and a hazard was identified for each of them (Table 5.7). For each hazard, a hazardous event was identified which was created in the simulation environment and experienced by the participants during each of the driving scenarios in the experiment runs. Similar to study one, the choice of hazardous events was also influenced by the ability to create the events in the simulation environment of 3xD simulator.

Table 5.7: Hazard and hazardous event description for study six

Function	Hazard	Hazardous event description
Object Detection (Camera + computation power)	Late detection of obstacle w.r.t. speed and computation power	Cyclist / Bike crosses a junction when he didn't have the right of way.
Object Detection (Radar)	False detection at a curve in foggy weather	Vehicle applying brakes when facing a curve as curve is misinterpreted as an obstacle in foggy weather
Computation power	Detection of obstacle w.r.t. speed and computation power	Another vehicle cuts-in into the lane of the ego vehicle at high speed on a motorway
Object detection (Camera)	Compromised detection due to environmental factors	In heavy rainy weather, vehicle is not able to detect traffic light or speed sign.
Object Detection (Camera + computation power)	Late detection of obstacle (as obstacle initially hidden behind other static objects)	Pedestrian (child) jumps out on road suddenly.

5.4.1.4.2. Automation capability

Similar to study one (trust and static knowledge: section 5.3), two levels of automation capability were used in study six (trust and dynamic knowledge): 1) low capability automation 2) high capability automation. The difference between the two systems was based on the ability of the automated system to tackle the five hazardous events mentioned in section 5.4.1.4.1. Similar to the system capability in study one (section 5.3.1.3.2), low capability automated system was able to handle one out of the five hazardous events (Table

5.8), requiring the driver to transition to manual mode in four hazardous events to prevent an accident or traffic rule violation. High capability automated system was able to handle four out of the five hazardous events, requiring the driver to transition to manual mode in only one hazardous event situation to ensure safe performance.

Table 5.8: Hazardous situations encountered by the autonomous vehicle during the study run

Function	**Hazard**	**Low Capability Automation**			**High Capability Automation**	
		Run 1	Run 2	Run 3	Run 1	Run 2
Object Detection (Camera + computation power)	Late detection of obstacle w.r.t. speed and computation power	Yes	No	Yes	Yes	Yes
Object Detection (Radar)	False detection at a curve in foggy weather	No	No	No	Yes	Yes
Computation power	Detection of obstacle w.r.t. speed and computation power	No	No	No	Yes	Yes
Object detection (Camera)	Compromised detection due to environmental factors	No	No	No	No	No
Object Detection (Camera + computation power)	Late detection of obstacle (as obstacle initially hidden behind other static objects)	No	Yes	No	Yes	Yes

5.4.1.5. Procedure

Similar to the procedure in study one (section 5.3.1.4), when participants arrived for the experiment, they were initially briefed about the experiment and informed consent and demography data was taken from each participant. In order to familiarize the participants with the simulator and the engagement/disengagement of automation, each participant was given a trial run on the driving simulator with the author seated next to the participant. Contrary to study one, participants sat in the driver's seat and were able to adjust the seat and mirror positioning as per their comfort needs. The simulator was configured in SAE Level 3 mode. The trial run was on a prebuilt scenario that was visually similar to the study run, but replicated none of the hazardous events. Each trial run was of five minutes duration. Participants were told that they can ask for as many trial runs as they wish in order to make themselves comfortable with the simulator environment. After the trial runs, participants were asked whether they would like to continue the study. In the case that the participant agreed, participants in group L1a and L1b experienced three experiment runs of around 15 minutes each, while participants in group L2 and group H experienced two experiment runs of around 15 minutes each. Participants were randomly assigned to the groups. Before the second run in all the groups, participants were informed that they would receive dynamic

knowledge about the automation system health via an HMI display. All participants experienced an additional demo run before the second run to familiarize themselves with the working of the HMI display and the scoring system. In order to examine the effects that static knowledge might have on trust ratings when provided along with real-time knowledge, group L1b received static knowledge about the capabilities of the automated system prior to the trial, in addition to the dynamic knowledge presented on the HMI display. This static knowledge was presented to the participants via a commentary that was read out to them via a prepared script, previously described in section 5.4.1.5.1.

At the end of each experiment run, participants were asked to fill a trust rating questionnaire (discussed in chapter 4), Simulator Sickness Questionnaire (SSQ) (Kennedy et al., 1993), and NASA Task Load Index (TLX) questionnaire (Hart and Staveland, 1988). The NASA TLX was administered via the official NASA TLX App for an iPad (IOS) developed by NASA.

5.4.1.5.1. Imparting Dynamic Knowledge

Dynamic knowledge was imparted to the participants via a customised HMI display. The content of the dynamic knowledge was influenced by the qualitative comments received from the participants of study one (discussed in section 5.3.3). In order to meet the comments like *"warnings from the car missing"* and *"no warnings & notification"*, the HMI display provided information about the automation system health which had three states:

- Green status: meant that the automated system is confident about its abilities to perceive the environment and to carry out the driving task.
- Amber status: meant that the automated system can carry out the driving task however it is unsure about the sensing capabilities in the environment as the system is at its operational boundaries.
- Red status: meant that the system is performing the driving task but doesn't take any responsibility to ensure safety of the system as the system is working outside its operational capabilities.

In order to instil knowledge-based behaviour, the HMI display also provided the reason for any change in the automation system health state to help participants calibrate their mental models. Providing the reason for the change in the automation system health was an important aspect of the imparted knowledge as it facilitated the participant's understanding of the working of the system and ensured that they didn't consider automation state changes as rules. In addition, the HMI display provided the participants with their real-time score,

any bonus/penalty received by them and the reason behind the bonus/penalty. Figure 5.12 illustrates various modes of the HMI design's display. Static knowledge was imparted via a prepared scripts as described in section 5.4.1.5.1. In manual mode, the automation health status was greyed out along with the text "Manual mode" being displayed on the HMI display, suggesting that the system was in manual mode.

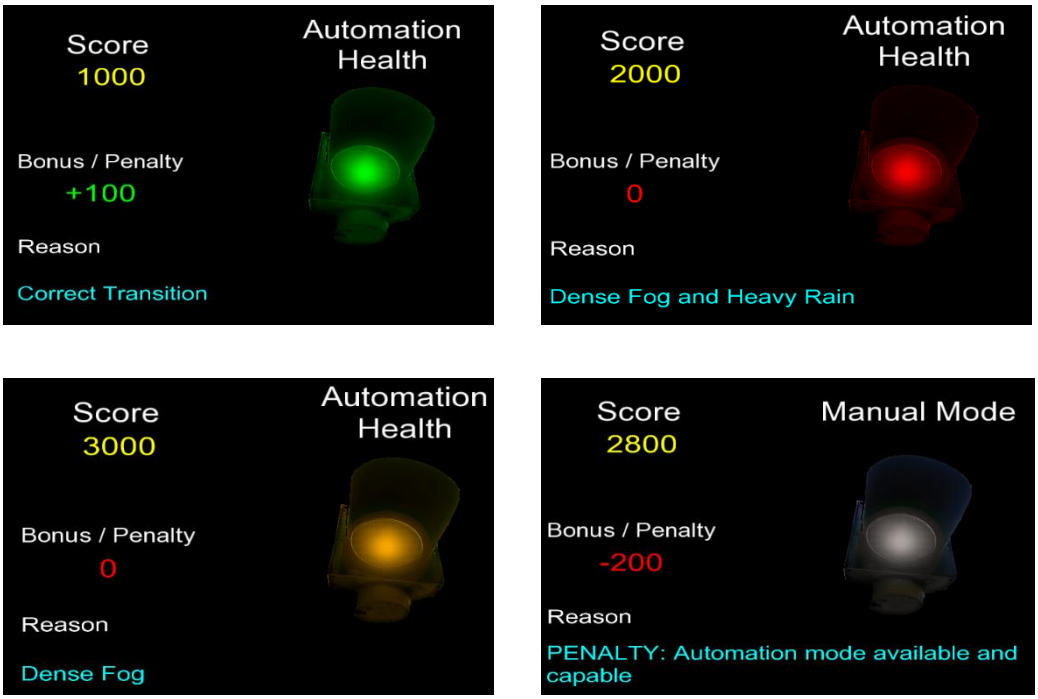

Figure 5.12: Various HMI display messages

In automated mode, whenever the automation health status changed, the HMI display was updated along with an auditory feedback message e.g. *"automation health is amber"*. A second feedback modality was added in order to avoid the situation of the driver failing to notice the visual change in automation health. Multimodal feedback has been shown to have better driver performance (Biondi et al., 2017; Jakus et al., 2015; Wickens, 2008). While in manual mode, in case automation was not switched on for more than 10 seconds after automation was available, participants were prompted to engage automation via an audio message: *"you may engage automation"*. This was done to ensure the objective of the experiment was met and to prevent fully manual drive by the participants. However, participants didn't receive any bonus after engaging automation due to the prompt and it was excluded from the counts for correct transitions. Additionally, participant scores were updated in real-time. One of the important aspects of the design of the HMI feedback was to provide the reason behind any status change to provide the participants with the context of the situation, enabling them to display knowledge based behaviour.

5.4.2. Results

5.4.2.1. Trust levels

As discussed in section 5.4.1.5.2, trust ratings were measured using a subjective rating scale and participants were asked to draw a line across a 100 mm box to indicate their level of trust (c.f. (Muir and Moray, 1996; Rajaonah et al., 2006)). The average *"trust in the system"* for low-capability automation increased (which was statistically significant) with the introduction of dynamic knowledge about the automation system health from 37.28% (S.D. = 26.34%) to 60.89%. (S.D. = 23.18%) (Figure 5.13). Participants experiencing high capability automation also showed a similar trend for *"trust in the system"* which increased (which was statistically significant) from 54.03% (S.D. = 31.39%) to 80.89% (S.D. = 14.96%) (Figure 5.13). In contrast to the findings of study one (section 5.3.2.1), it is interesting to note that with dynamic knowledge, both low and high-capability automation systems had relatively similar increase in trust ratings leading to different mean and median values for the run with dynamic knowledge (Figure 5.13). In both low-capability automation and high-capability automation group barring one participant, remaining 13 participants (in each group) showed an increase in trust-in the system ratings with the introduction of dynamic knowledge (Figure 5.14).

A repeated measures ANOVA was conducted for the *"trust in the system"* ratings with automation capability as a between factor variable and dynamic knowledge as the within factor variable. The introduction of dynamic knowledge about the automation system health and reasons behind it had a highly significant statistical effect on *"trust in the system"* ratings, $F(1, 26) = 37.99$, $p = 0.000002$ with a $\eta_p^2 = 0.594$, suggesting 59.4% of the variance being associated with the introduction of dynamic knowledge via the HMI display. There was no interaction effect between the factors automation capability and dynamic knowledge for *"trust in the system"* ratings, $F(1, 26) = 0.16$, $p = 0.695$ ($\eta_p^2 = 0.006$).

The average *"trust with the system"* for low-capability automation increased with the introduction of dynamic knowledge from 48.14% (S.D. = 28.79) to 66.42% (S.D. = 29.53) (Figure 5.15). Similar trend was observed for the high-capability automation group in which *"trust with the system"* increased from 44.92% (S.D. = 33.27%) to 76.25% (S.D. = 19.76%), with the introduction of dynamic knowledge (Figure 5.15). In the low-capability automation group, similar to the trust-in the system ratings, barring one participant, remaining 13 participants showed an increase in trust-with the system ratings with the introduction of the dynamic knowledge. A repeated measures ANOVA was conducted for the *"trust with the*

system" ratings with automation capability as a between factor variable and dynamic knowledge as the within factor variable. Providing dynamic knowledge via the HMI display had a statistically significant effect on *"trust with the system"*, F (1, 26) = 24.91, p = 0.000034 with a $\eta_p^2 = 0.489$. There was no interaction effect between the factors automation capability and dynamic knowledge for *"trust with the system"* ratings, F (1, 26) = 1.72, p = 0.201 ($\eta_p^2 = 0.062$).

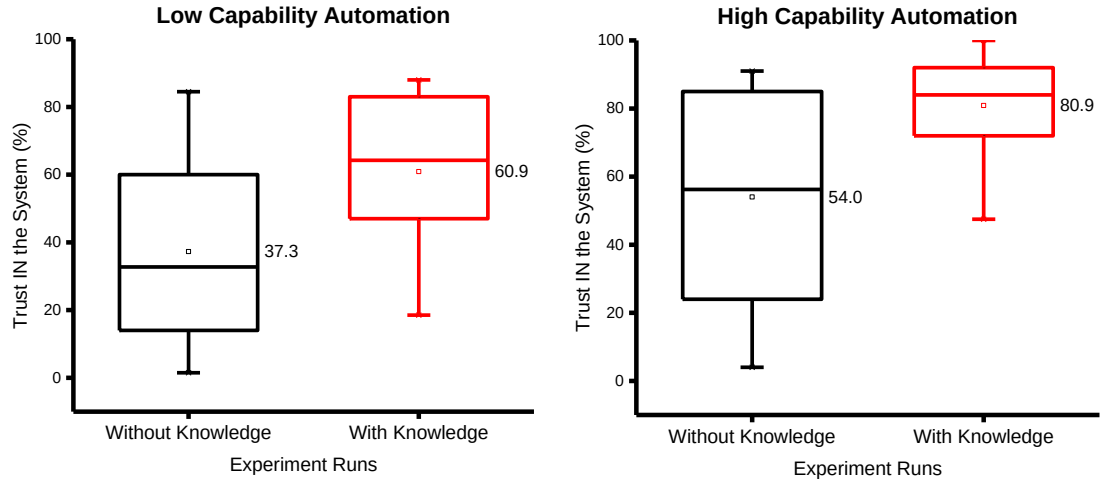

Figure 5.13: Box-plots of Trust-In the system ratings (highlighting average trust ratings) (central dot represents average value)

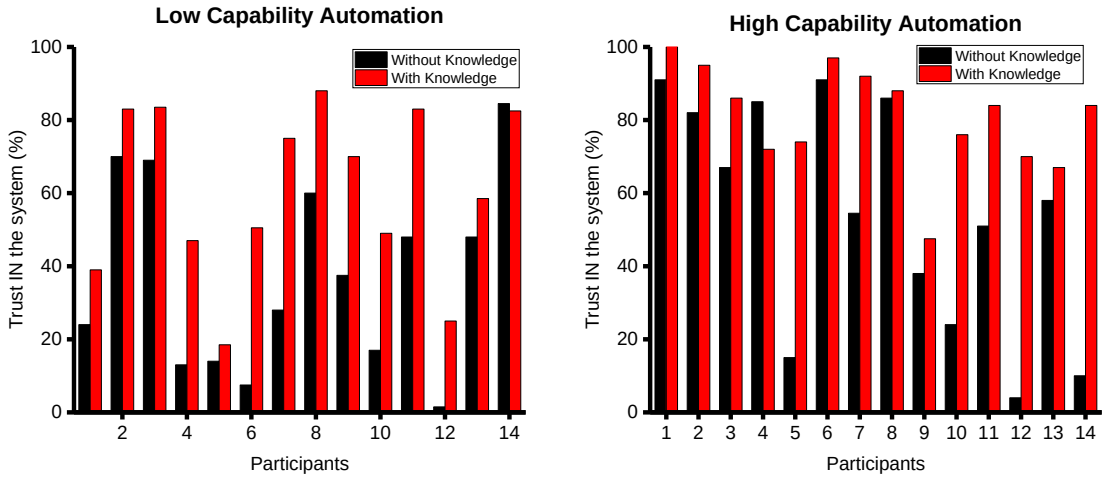

Figure 5.14: Trust IN the system ratings (study six: effect of dynamic knowledge on trust)

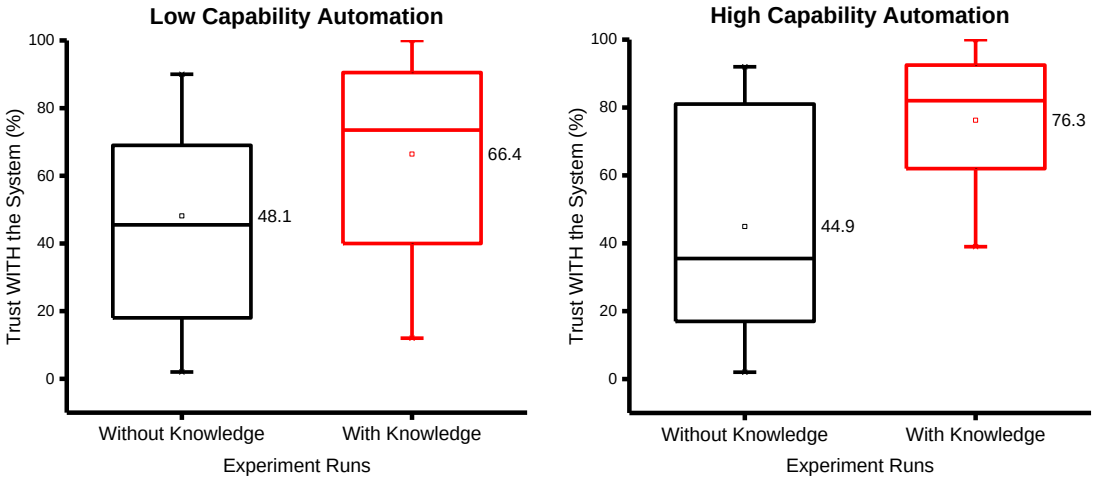

Figure 5.15: Box-plots of Trust-WITH the system ratings (highlighting average trust ratings) (central dot represents average value)

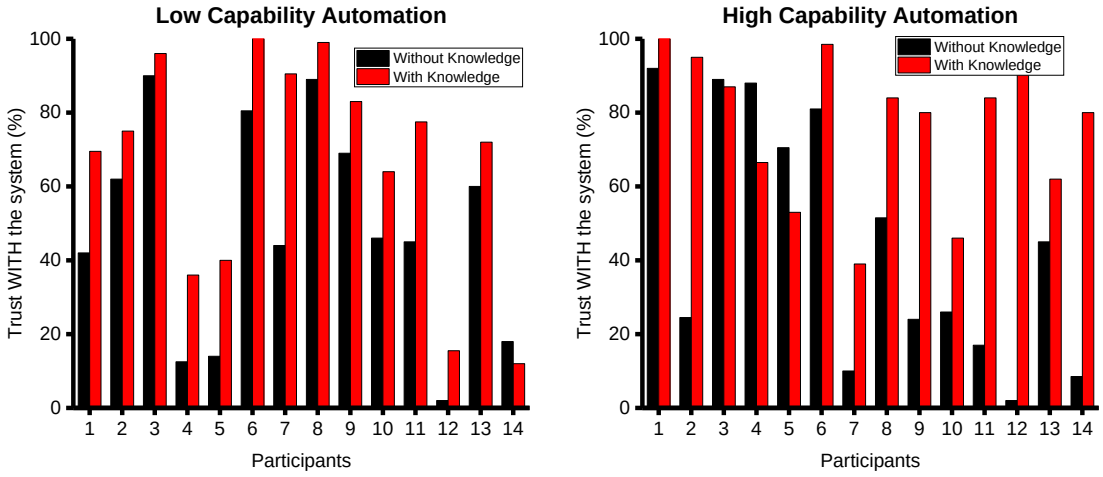

Figure 5.16: Trust WITH the system ratings (study six: effect of dynamic knowledge on trust)

5.4.1.5.1. Groups with three study runs (Group L1a and L1b): Evaluating learning effect

In order to negate the possible argument that increase in trust levels occurred due to a learning experience irrespective of HMI display and knowledge type, groups L1a and L1b were designed in the study. In both the groups, participants experienced three runs with no (static or dynamic) knowledge being provided in run 1 and run 3 (first and last run). In run 2 (second run), dynamic knowledge was provided to participants via HMI display in group L1a and dynamic knowledge along with static knowledge was provided in group L1b. The average ratings for *"trust with the system"* and *"trust in the system"* for group L1a and L1b are summarized in Table 5.9 and Table 5.10 respectively (and Figure 5.17 and Figure 5.18).

Table 5.9: Average ratings for Trust WITH the system for groups L1a and L1b

Participant Group	Trust WITH the System (out of 100)		
	Run 1 (no knowledge)	**Run 2 (with knowledge)**	**Run 3 (no knowledge)**
Group L1a	50.17 (S.D. = 32.98)	69.41 (S.D. = 27.03)	36.75 (S.D. = 24.45)
Group L1b	41.58 (S.D. = 26.21)	77.58 (S.D. = 23.07)	34.67 (S.D. = 22.50)

Table 5.10: Average ratings for Trust IN the system for groups L1a and L1b

Participant Group	Trust IN the System (out of 100)		
	Run 1 (no knowledge)	**Run 2 (with knowledge)**	**Run 3 (no knowledge)**
Group L1a	32.91 (S.D. = 28.83)	53.58 (S.D. = 25.52)	30.75 (S.D. = 19.37)
Group L1b	43.33 (S.D. = 33.02)	66.58 (S.D. = 27.41)	27.5 (S.D. = 14.12)

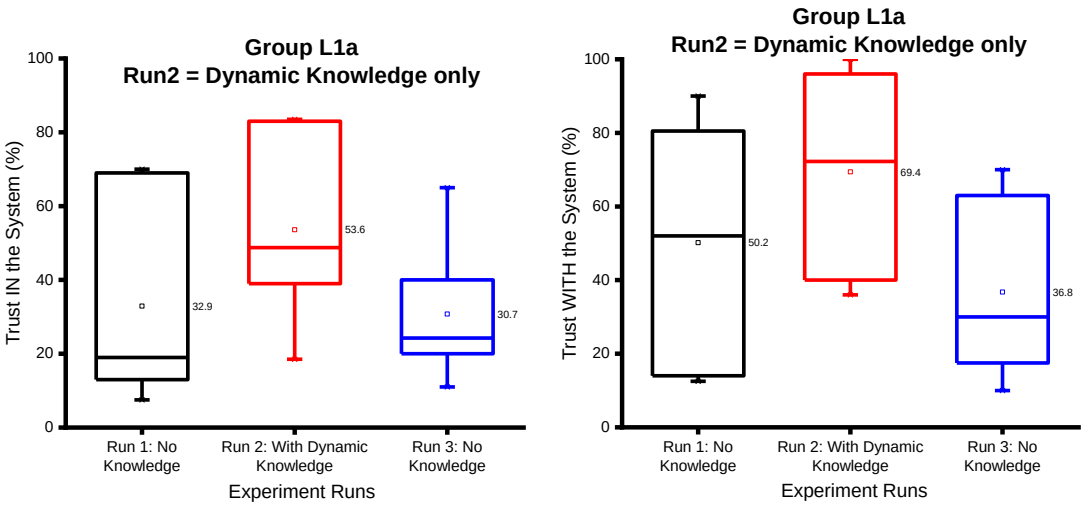

Figure 5.17: Box-plots of Trust ratings for group L1a (highlighting average trust ratings) (central dot represents average value)

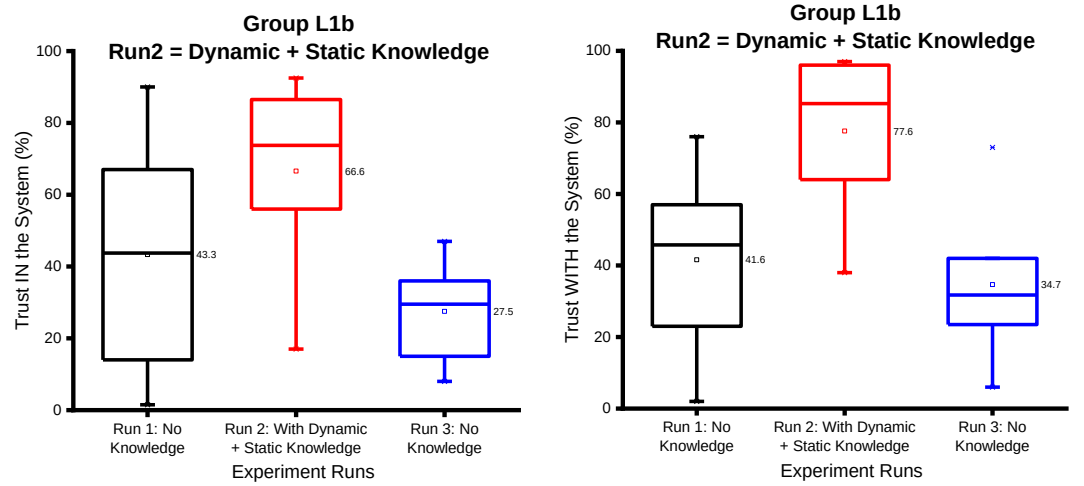

Figure 5.18: Box-plots of Trust ratings for group L1b (highlighting average trust ratings) (central dot represents average value)

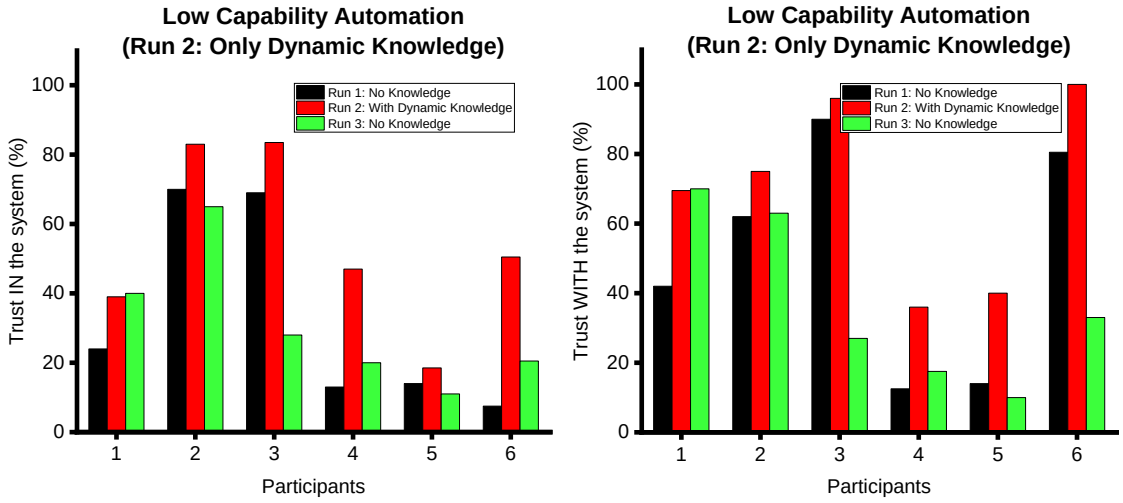

Figure 5.19: Trust ratings for groups L1a (run 2: only dynamic knowledge)

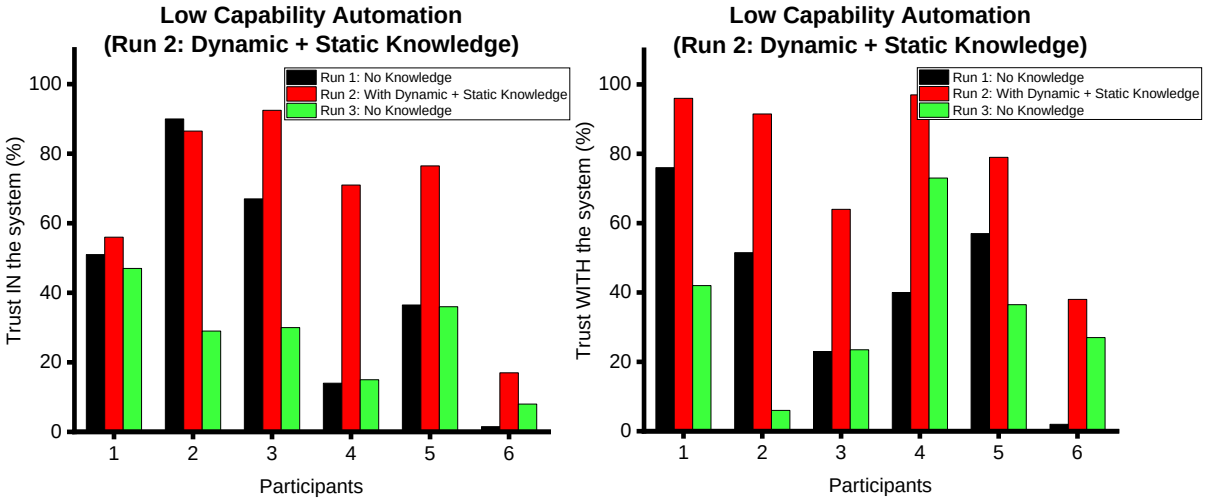

Figure 5.20: Trust ratings for groups L1b (run 2: both dynamic and static knowledge)

It is interesting to note that the trust ratings in run 2 for both *"trust in the system"* and *"trust with the system"* were higher for group L1b as compared to group L1a, suggesting the combined effect of dynamic and static knowledge is higher than only dynamic knowledge. This further supports the finding discussed in study one (section 5.3) that static knowledge has a significant main effect on *"trust in the system"* rating.

For both *"trust with the system"* and *"trust in the system"* ratings, trust level increased with the introduction of knowledge (in run 2) and decreased again in run 3, when no knowledge was provided. A pair-wise repeated measures ANOVA showed a significant main effect of the introduction of knowledge and no effect (insignificant) due to experience. The results of the ANOVA have been summarized in Table 5.11.

Table 5.11: Repeated Measures ANOVA on Trust ratings for L1a and L1b groups

Trust type	Participant Group	Pair	"p" value
Trust WITH the system	L1a	Run 1 – Run 2	0.002
		Run 2 – Run 3	0.04
		Run 1 – Run 3	**0.385**
	L1b	Run 1 – Run 2	0.001
		Run 2 – Run 3	0.01
		Run 1 – Run 3	**0.618**
Trust IN the system	L1a	Run 1 – Run 2	0.018
		Run 2 – Run 3	0.037
		Run 1 – Run 3	**0.809**
	L1b	Run 1 – Run 2	0.053
		Run 2 – Run 3	0.011
		Run 1 – Run 3	**0.210**

5.4.2.2. Degree of Incorrect trust, Degree of Appropriate trust and the Trust Continuum

While subjective ratings for trust have been discussed in section 5.4.2.1, it could be argued that they are prone to individual variations (even though statistical results suggest otherwise). Additionally, subjective rating scales are retrospective in nature (i.e. participants rate their experience after they have experienced the automated system). In this section, two concepts to objectively evaluate quality of trust in automation are introduced. These are Degree of Incorrect Trust (DoIT) and Degree of Appropriate Trust (DoAT). Both concepts build on the literature on *distrust*, *mistrust* and *appropriate trust* discussed in chapter 3. In the context of this study, distrust and mistrust could be evaluated in terms of the number of mismatching or matching interventions with respect to the number of correct human interventions required for safe completion of the experiment route.

DoIT represents the driver's unwillingness to trust the capability of the automated system leading to unnecessary interventions to disengage the automated system when it could have coped with a situation or to engage the automated system when it cannot cope with a situation. DoIT can have a value between 0 and 1. Higher the DoIT value, higher the incorrect trust of the driver.

$$Degree\ of\ \text{Incorrect Trust (DoIT)} = \frac{f}{1+f}$$

Where,
f = number of false transitions

The average values for Degree of Incorrect Trust (DoIT) have been summarized in Table 5.12. While DoIT showed an increase for both low capability and high capability automation groups, none of the increases was statistically significant. However, DoIT increase with knowledge for high capability automation was nearly significant. A repeated measures ANOVA for DoIT with dynamic knowledge as the within subject factor showed that there was no significant difference with dynamic knowledge on DoIT for low capability automation group, $F(1, 13) = 0.008$, $p = 0.538$ and for high capability automation group, $F(1,13) = 4.461$, $p = 0.055$.

Table 5.12: Average Degree of Incorrect Trust (DoIT)

Group	**Average Degree of Incorrect Trust (DoIT)**	
	Without knowledge	**With knowledge**
Low capability automation	0.70	0.73
High capability automation	0.487	0.697

DoAT represents the appropriateness of the driver's trust level, i.e, driver's correct trust (trust in right situations). DoAT's formulation implicitly incorporates over trust within its calculation. Over trust would result in a missed intervention which would mean that the number of correct interventions would be decreased.

$$Degree\ of\ Appropriate\ Trust\ (DoAT) = \frac{c - f}{c_{required}}$$

Where,
c = number of correct transitions
f = number of false transitions
$c_{required}$ = number of required correct transitions

The average values for Degree of Appropriate Trust (DoAT) have been summarized in Table 5.13. With the introduction of dynamic knowledge, DoAT showed an increase for both low capability and high capability automation groups. For high capability automation increase in DoAT was statistically significant while for low capability automation the increase for nearly significant. A repeated measures ANOVA for DoAT with knowledge as the within subject variable for low capability automation group resulted in in no significant difference, $F(1,13) = 3.976$, $p = 0.068$ and for high capability automation group resulted in, $F(1,13) = 8.163$, $p = 0.013$.

Table 5.13: Average Degree of Appropriate Trust (DoAT) values

Group	Average Degree of Appropriate Trust (DoAT)	
	Without knowledge	With knowledge
Low capability automation	0.330	0.508
High capability automation	0	0.321

One of the participant's data (participant # 4) was found to be an outlier (using boxplots) in various trust ratings and was removed from the analysis.

5.4.2.3. Time spent in correct mode

The introduction of dynamic knowledge had a positive effect on the percentage of time spent by the participants in the correct mode. For high capability automation, the average percentage of time spent in the correct mode for the vehicle showed an increase from 82.39% (S.D. = 4.80) to 91.96% (S.D. = 2.39) with the introduction of dynamic knowledge via the HMI display. A similar trend was observed for low capability automation, where the average time spent in correct mode by the participants increased from 79.05% (S.D. = 6.93%) to 84.50% (S.D. = 4.77%) with the introduction of dynamic knowledge.

A repeated measures ANOVA with dynamic knowledge as within subject variable showed a highly statistically significant main effect of the introduction of dynamic knowledge on percentage of time spent in correct mode, $F(1,13) = 49.363$, $p = 0.000009$ with a $\eta_p^2 = 0.792$, suggesting 79.2% of the variance occurring due to the introduction of dynamic knowledge in run 2. Similar results were obtained for the low capability automation group, where a repeated measures ANOVA with dynamic knowledge as a within subject variable showed a significant main effect on the percentage of time spent in correct mode, $F(1,13) = 11.227$, $p = 0.005$ with a $\eta_p^2 = 0.463$.

5.4.2.4. Workload

The average workload, as measured by the NASA TLX subjective questionnaire, for low capability automation showed a statistically insignificant change from 51.49 (S.D. = 12.08) to 46.79 (S.D. = 15.05) with the introduction of dynamic knowledge. Similarly, for high capability automation, the introduction of dynamic knowledge showed a statistically insignificant change from 43.27 (S.D. = 9.73) to 46.67 (S.D. = 16.03). A repeated measures ANOVA with dynamic knowledge as within subject variable showed a statistically insignificant main effect of the introduction of dynamic knowledge on workload for both low capability automation, $F(1,12) = 1.856$, $p = 0.198$ ($\eta_p^2 = 0.134$); and high capability automation, $F(1,13) = 0.875$, $p = 0.367$ ($\eta_p^2 = 0.063$).

This is in line with the hypothesis that dynamic knowledge will not have a significant effect on workload. Furthermore, dynamic knowledge (if designed optimally) has a potential to reduce workload, or at the very least not increase it, while offering the benefits of increased appropriate trust as shown in section 5.4.2.1 and 5.4.2.2. This is evident in the workload ratings where most participants (9 out of 12 in low capability automation and 6 out 14 in high capability automation) reported a decrease in workload with the introduction of dynamic knowledge for both low capability and high capability automation (Figure 5.21). This finding is in contrast to the finding in section 5.3.2.4, where introduction of static knowledge significantly increased workload experienced by the drivers ($p = 0.003$). However, when static knowledge was provided in conjunction with dynamic knowledge (run 2 in group L1b), a non-significant decrease was observed in the workload ratings as compared to run 1 where no knowledge was provided, $F(1, 5) = 2.342$, $p = 0.186$ ($\eta_p^2 = 0.319$).

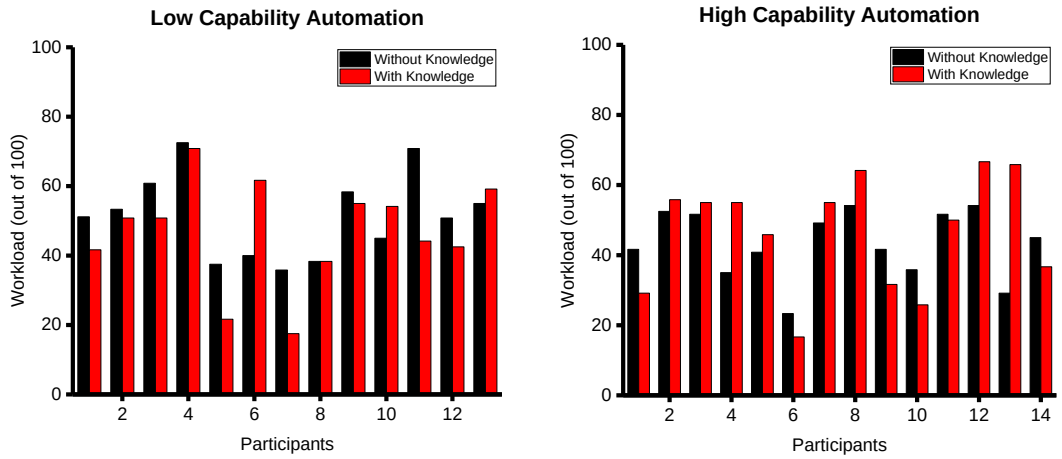

Figure 5.21: Workload ratings (NASA TLX) for various participants

5.4.3. Discussion

This simulator study (study six: trust and dynamic knowledge) demonstrated the effect of dynamic knowledge of capabilities and limitations of the system on trust. The results show a statistically significant increase in both *"trust in the system"* and *"trust with the system"* ratings with the introduction of dynamic knowledge. Additionally, Degree of Appropriate Trust (DoAT) also significantly increased with introduction of dynamic knowledge, suggesting increased level of "correct" trust leading to correct use of the automated systems by the drivers.

It is interesting to note that the introduction of dynamic knowledge had a similar effect for both low capability and high capability automation groups leading to different mean and median values for trust ratings in the two groups with the introduction of dynamic knowledge. In contrast, when only static knowledge was provided (in study one on trust and static knowledge: section 5.3), trust in the system ratings for both high and low capability automation were at similar levels (Figure 5.6). This suggests that dynamic knowledge potentially has a greater impact on trust ratings as compared to static knowledge. Some of the qualitative feedback from the participants support this inference. One of the participants (participant #5) commented, *"I know when I have to be more alert"*. A similar comment was made by another participant (participant #10) who commented *"visual information (red, amber or green) provides a measure of how alert driver (I) should be"*.

In order to better understand the increase in workload reported by some of the participants in the high capability automation group, the author will discuss the qualitative feedback from the participants which explains their ratings. One of the participants who reported higher workload with the introduction of dynamic knowledge commented that the reason behind the increase was the position of the HMI display. The participant reported that he would expect the HMI display to be in front of the driver to avoid looking towards the display intermittently. While the comment was perfectly valid, study six was not designed to evaluate the effect of positioning of the HMI display. Moreover, only one of the participants (out of 35 participants) made such a comment. Another participant who reported higher workload with the introduction of dynamic knowledge (participant # 26) reported that their workload would decrease if they had more experience with the system (HMI display). The participant commented *"I would prefer system 2 (run 2) but more time is needed to understand its limits because more information is given to the driver"*.

It is interesting to note that for high capability automation, both mean and median for *"trust in the system"* was higher than *"trust with the system"* in both no knowledge and with dynamic knowledge runs, while reverse was true for the low capability automation group. This can potentially be explained by some of the qualitative feedback received from the participants of the high capability automation group. One of the participants (participant # 25) who experienced high capability system suggested the increase in workload and lower ability to work with the system when dynamic knowledge was introduced, was due to the difference in the nature of the workload required in run one and run two. According to the participant, *"system 1 (run 1) (required) an effort to avoid switching off mentally"* while *"system 2 (run 2) (required) an effort to keep concentrating all the time for its limitations"*. High capability automation kept the drivers out-of-the-loop, i.e. driving task, for a longer duration as the participants were required to make only one intervention. In contrast, for low

capability automation, participants were more involved in working with the system and therefore didn't experience out-of-the-loop state for long durations. While the aim of manufacturers should always be to introduce automated systems with the highest possible capabilities, this is not always possible due to limitations in current sensing capabilities and lack of real-world testing. However, in order to have high trust with the system, it is important to have the driver engaged with the system either physically or in an interactive manner (e.g. intelligent tutoring (Beggiato et al., 2015; Beggiato and Krems, 2013)).

5.5. Informed Safety

The author introduces the concept of *"informed safety"*, as a means to increase trust on automated systems and calibrate it to appropriate levels. Informed safety means informing the driver (via static and dynamic knowledge) about the safety limits of the automated system and its intention. Foundations of informed safety lie in the Rasmussen's Skills-Rules-Knowledge (SRK) model and the abstraction hierarchy (discussed in section 5.1.1 and chapter three). Inspired by the SRK model, moving decision making by the drivers to a knowledge based level, enables the driven to understand the context of the decision ensuring correct use even in situations they have not experienced before.

From the results of study one (discussed in section 5.3), it could be inferred that vehicle manufacturers may choose to introduce low-capability systems and provide static knowledge in order to deliver increased user trust and overall system performance. However, there is a caveat to this inference. For low capability automation, while introduction of static knowledge increased the level of *"trust in the system"* significantly (from 32.4% to 65.4%) and had a statistically significant main effect ($p = 0.000002$), it also increased the number of false presses significantly (from 0.476 to 2.67) and had a statistically significant main effect on workload (measured using OWS), suggesting an increase with the introduction of static knowledge ($p = 0.003$). Therefore, very low capability and too much knowledge is also not an appropriate solution. The author believes that there is an optimum level of system capability and knowledge to be imparted at which trust could be maximized, and false presses and workload could be minimized. Therefore, manufacturers may decide to enhance automation capability by providing knowledge. Until systems are fully (100%) capable, augmenting system capability with knowledge about the system's true capabilities, could be a method to bridge the gap in trust. In other words, while manufacturers should aim to introduce high capability systems in the market, the gap in system capability (system limitations) should be provided as knowledge to the customers to ensure high and appropriate trust in the system. While providing dynamic knowledge could help reduce

workload, further research is required to establish the optimum level of system capability and knowledge to be provided to the driver.

It is a well-known fact that users don't read manuals and that vehicle dealers/Original Equipment Manufacturers (OEMs) rarely do a good job in sufficiently or appropriately informing customers about the system capabilities and limitations (Beggiato and Krems, 2013; Eichelberger and McCartt, 2014; Larsson et al., 2014). As automated systems are introduced, innovative methods of informing the driver (customer) to create an *"informed safety"* level, need to be implemented. One potential solution could be providing a virtual tour of the vehicle at the dealership, which gives the customers an immersive experience of the various features and can help them calibrate their mental models and their expectations from the vehicle. Other means from providing *"informed safety"* may be short videos on the working of the Human Machine Interface (HMI) or specifically designed voice assistant features. All the discussed methods may form a part of the initial showroom briefing or a pre-sale briefing. However, these methods need to be evaluated to measure their effectiveness.

It is important to appreciate the difference in the manner in which non-specialists (i.e. general public) would understand / interpret the knowledge imparted to them. As creators of the system, designers and engineers have an appreciation and inclination towards technical understanding and the technical feature explanation. Therefore, in study one (section 5.3) care was taken in the language used in the script used to impart knowledge to the participants. Use of technical jargon terms was avoided and illustrations were used as examples to help participants visualize the system. In real life, it is important that manufacturers explain the system capabilities and limitations in a non-technical manner in order to aid customer's understanding by providing examples and ensuring the people read the provided information.

In addition, study six (discussed in section 5.4) has shown that dynamic knowledge about the automation system capabilities and limitations is able to overcome some of the shortcomings of just providing static knowledge about the system. While introduction of dynamic knowledge had a significant main effect in the increase of the *"trust in the system"* and *"trust with the system"* ratings, contrary to the effect of static knowledge, it didn't have significant effect on the number of false presses for both the low capability and high capability automation ($p = 0.902$ and $p = 0.519$ respectively). Introduction of dynamic knowledge has shown the potential to reduce the workload of the drivers too.

Thus, the shortcomings of providing static knowledge can be overcome by augmenting it with dynamic knowledge about the capabilities and limitations of the automated system.

Hence, in order to create an informed safety level for customers, both static and dynamic knowledge needs to be provided. Therefore, manufacturers may decide to enhance automation capability by providing static and dynamic knowledge. However, until systems are fully (100%) capable, augmenting system capability with knowledge about the system's true capabilities, could be a method to bridge the gap in trust. In other words, while manufacturers should aim to introduce high capability systems in the market, the gap in system capability (system limitations) should be provided as knowledge to the customers developing their informed safety level to ensure high trust in the system.

This research introduces the concept of *"informed safety"*, as a means to calibrate trust to the appropriate levels, which may include increasing those with low trust in capabilities or even reducing trust in those with too much trust in what the system can achieve by making them aware of system boundaries. Informed safety means informing the driver (via static and/or dynamic knowledge) about the safety limits of the automated system and its intention. Informed safety provides the ability to display knowledge-based behaviour to shift the interpretation of a scenario to higher abstraction level or a lower abstraction level (Rasmussen, 1983). Informed safety aids the driver to interpret an unexpected situation to adopt an appropriate tactical or strategic manoeuvre to handle the situation safely. Informed safety is not just about providing rules of usage, it includes the background information, understanding and knowledge about how the system operates.

5.6. Conclusion

Study one (static knowledge) and study six (dynamic knowledge) have shown that trust on automated systems can be increased with increasing drivers' informed safety by providing them static and dynamic knowledge about the automated system's true capabilities and limitations. While static knowledge significantly increased *"trust in the system"* ratings ($p = 0.000002$), dynamic knowledge had a significant increase on both *"trust in the system"* ($p = 0.000002$) and *"trust with the system"* ($p = 0.000034$). However, when both static and dynamic knowledge is provided to the driver, the increase in trust is much higher than when only dynamic knowledge is provided. It is interesting to note that static knowledge caused a significant increase the workload levels in a low capability automation system. In contrast, dynamic knowledge had no such main effect on workload for either low capability or high capability automation. Additionally, when both static and dynamic knowledge was provided, the change in workload was statistically insignificant when compared to no knowledge.

While the two studies have illustrated the potential benefits of providing static and dynamic knowledge to create an informed safety level for the driver which in turn has the potential of

increasing *"trust in the system"* and *"trust with the system"*, the way to create the knowledge for informed safety needs to be explored. While dynamic knowledge can be imparted using an in-vehicle HMI display, some of the potential ways of imparting static knowledge could include a virtual tour of the vehicle at the dealership or short videos on the system working. These could be a part of the pre-sale briefing at the car showroom.

Having established the effect of knowledge (static and dynamic) on trust, the content of the knowledge is an important aspect of improving trust on automated systems. The two driving simulator studies in this chapter have shown that the importance of establishing the design limitations of the automated system and presenting these to the user in order to increase appropriate use of automated systems. Creating this knowledge requires a robust and a reliable verification and validation process which can identify and categorize various hazardous situations and test the systems in those situations. In the next chapters (chapter 6 and chapter 7), the author will explore the creation of various test scenarios to establish the knowledge of design limitations of the automated systems and reliably characterize them to provide create drivers' informed safety level.

AUTOMATED VEHICLES: DEMYSTIFYING THE CHALLENGES – PART 2 (TESTING)[3]

Chapter 6

Literature Review

In chapter two, the author identified various opportunities and challenges associated with Advanced Driver Assistance Systems (ADASs) and Automated Driving Systems (ADSs). However, in order to reap the benefits of the ADAS and ADS, it is essential that drivers use the systems. Therefore, it becomes imperative that the solutions are found to the challenges offered by ADAS and ADS.

In this chapter, one of the challenges is discussed in detail and subsequently associated research questions are identified, which will be tackled within the scope of this thesis. This chapter discusses the challenges associated with establishing the safety level of ADAS and ADS (which form the knowledge to be imparted to the drivers to create the state of informed safety).

[3] Contents of this chapter have been published in the following publications:

Khastgir, S., Dhadyalla, Gunwant, Birrell, S., Redmond, S., Addinall, R., et al. (2017) 'Test Scenario Generation for Driving Simulators Using Constrained Randomization Technique', in SAE Technical Paper# 2017-01-1672. doi: 10.4271/2017-01-1672.

Khastgir, S. et al. (2018b) 'The Science of Testing: An Automotive Perspective', in SAE Technical Paper: 2018-01-1070. doi: 10.4271/2018-01-1070.

Khastgir, S., Birrell, S., Dhadyalla, G., Sivencrona, H., et al. (2017) 'Towards increased reliability by objectification of Hazard Analysis and Risk Assessment (HARA) of automated automotive systems', Safety Science. Elsevier Ltd, 99, pp. 166–177. doi: 10.1016/j.ssci.2017.03.024.

Khastgir, S., Sivencrona, H., et al. (2017) 'Introducing ASIL inspired Dynamic Tactical Safety Decision Framework for Automated Vehicles', in Proc. of the IEEE Conference on Intelligent Transportation Systems, Proceedings (ITSC) 2017. Yokohama. doi: 10.1109/ITSC.2017.8317868.

6.1. Part 2: Testing

Chapter three proposes and chapter five demonstrates (via driving simulator studies) that knowledge can potentially be used as an intervention method to calibrate drivers' trust to "correct" level to ensure safe use of the ADAS and ADS. To provide this knowledge, first it needs to be created and then imparted to the drivers to potentially enable them to have appropriate trust (discussed in chapter 5). It further discusses three types of knowledge based on their type of delivery: 1) static, 2) dynamic 3) internal mental model. This chapter focusses two aspects of knowledge that are common to each of the knowledge delivery types. These two aspects are:

- Content of knowledge
- Reliability of the content of knowledge

Testing and certification are potential ways of creating the *"knowledge"* under discussion. It is equally important that the process of creation of knowledge is reliable, such that it generates the same results when the process is carried by different testers and at different points in time. Thus, this chapter discusses the process of creation of *"reliable knowledge"* that could be imparted to the drivers to increase their trust level and at the same time, calibrate it to the correct level.

Before discussing the testing and certification process, it is important to define the two terms *"testing"* and *"certification"*. In order to define *"testing"*, the author adopts the definition from the international standard ISO 29119 – 1 (ISO, 2013) as *"set of activities conducted to facilitate discovery and/or evaluation of properties of one or more test items"*. While *"testing"* is defined in another international standard ISO 26262 – 2018 as *"process of planning, preparing, and operating or exercising an item… to verify that it satisfies specified requirements…"*, it falls short in capturing the whole space of testing since it refers specifically only to requirements based testing. Thus, the ISO 26262 – 2018's definition of testing has been discounted in this thesis. The composition of test space is discussed further in section 6.1.1.

While discussing certification as a factor influencing trust (chapter three), the author mentioned the adoption of the definition of *"certification"* as *"a written guarantee that a system or component complies with its specified requirements and is acceptable for operational use"*. Definition for certification can be seen to have two distinct aspects to it. First being *"complies with its specified requirements"* and second being *"acceptable for operational use"*. While the former aspect suggests that the system meets a desired defined performance, the later suggests that level of performance requirements is open for subjective

interpretation. These two concepts will be discussed in further detail in sections 6.1.1 and 6.1.2.

From the definitions of certification and testing, it can be inferred that testing is a way of achieving certification. Testing has three constituting elements:

- Test scenarios (discussed in section 6.1.1)
- Safety analysis (discussed in section 6.1.2)
- Test methods (overview in appendix two)

While new research is being carried out to develop new test methods ((Bock and Maurer, 2007; Huang et al., 2016; Ma et al., 2018; Rene et al., 2016; Schöner and Hurich, 2015)), detailed discussion on test methods remains out of scope of this thesis. An overview of the current methods and challenges is discussed in Appendix A2 in order to provide context to this research. However, test methods is not the focus of this research. While research on test methods improves on an aspect of existing test methods (e.g. improved representation of the environment, better reproducibility of tests, etc.), all test methods (discussed in appendix two) and the upcoming test methods have one important thing in common. All test methods need to identify test scenarios that they need to run using the developed test method. The next section (section 6.1.1), discusses test scenarios in greater detail, identifying the research gap and research questions associated with developing test scenarios.

6.1.1. Test Scenarios

6.1.1.1. Introduction

In the automotive industry, many relatively advanced features, have caused vehicle re-calls due to buggy software, costing millions of dollars to the manufacturers; e.g. to fix the ignition switch issue, General Motors spent approximately $400 million for the 2.6 million affected vehicles (Malek, 2017). Fixing a bug during the development process costs an average of $25, while after release it increases to $16000 on an average (Altinger et al., 2014). A bug in a released product could be caused due to: 1) release of untested code, 2) testing sequence differs from use sequence 3) user applied untested input values 4) untested operating conditions (Whittaker, 2000). The latter was illustrated in the Ariane 5 disaster (Lions, 1996), where software was reused from Ariane 4 software in the Ariane 5 system without enough testing (Weyuker, 1998). This importance of operating environment and potential consequence of untested inputs was also seen in the recent Tesla “Auto-pilot” system crash (NHTSA, 2017a). However, the Tesla “Auto-pilot” crash also had an improper human-computer interaction as a contributory factor. It has been suggested that majority of

the software related accidents are a result of the operation of the software rather than its lack of operation (Leveson, 2006).

Therefore, in order to ensure that the systems have a safe and a robust functionality, it is important to be able to define test scenarios which are able to: 1) trigger real-world use sequence 2) represent user input values 3) define and identify "all" operating conditions. However, lack of standardized methods test scenario definition or classification, and the lack of international standards to define safety requirements for ADASs and ADSs, have led to a subjective interpretation of test scenarios and desired "safety" levels, particularly for ADASs and ADSs in vehicles.

While the ISO 26262 standard (ISO, 2018b), provides some guidance for testing methods and approaches for a product development cycle, it too falls short to deal with the complexities of ADASs and ADSs. Furthermore, even with ISO 26262 been increasingly adopted in the industry, there is still a lack of a *"quantified and rigorous process for automotive certification"* (Yu et al., 2016). This is caused due to the lack of objective quantification of severity, exposure and controllability ratings which comprise the ASIL rating, causing inter- and intra-rater variation (Yu et al., 2016).

Current luxury cars are a complex system with over 100 million lines of code as compared to 6.5 million in a Boeing 787 airplane (Charette, 2009; Strandberg et al., 2018). The introduction of ADASs and ADSs is going to further increase the complexity manifold. While a variety of ADASs and ADSs exist or are in development, each of them offers a different kind of a challenge for testing. The move towards higher levels of automation is coupled with the challenge of testing and safety analysis as it needs complex solutions to include interactions between a larger number of variables and the environment. It is suggested that in order to prove that automated vehicles are safer than human drivers, they will need to be driven for more than 11 billion miles (Kalra and Paddock, 2016b). Even after 11 billion miles, such testing will *"only assure safety but not always ensure it"* (Transport Systems Catapult, 2017), thus suggesting vehicle level testing or real world testing before start of production (SOP) wouldn't be enough to prove safety of the automated driving systems (Koopman and Wagner, 2016; Wachenfeld and Winner, 2017a).

While software testing has been said to be the *"least understood part of the (system) development process"* (Whittaker, 2000), the author believes that a scientific approach needs to be adopted to solve the challenge of identifying scenarios that capture the complex interactions within systems and system-environment in an efficient manner.

6.1.1.2. Understanding scenarios

Scenarios as a term has been widely used as a buzzword because of its vague interpretation and myriad of uses (Campbell, 1992). Before delving in deep discussion about scenarios, it is important to understand the meaning of use cases, scenarios and test cases in context of this thesis.

A use case describes the system behaviour as a sequence of actions linking the result to a particular actor. A test scenario is a specific path through a use case, i.e., a specific sequence of actions. As illustrated in Figure 6.1, a use case can have many test scenarios which represent the system behaviour. Each test scenario can in turn have multiple test cases. A test case is a set of test case preconditions, inputs (including actions, where applicable), and expected results, developed to drive the execution of a test item to meet test objectives, including correct implementation, error identification, checking quality, and other valued information. Thus, a use case, test scenario and test case have a pyramid relationship, with the use case sitting at the top of the pyramid.

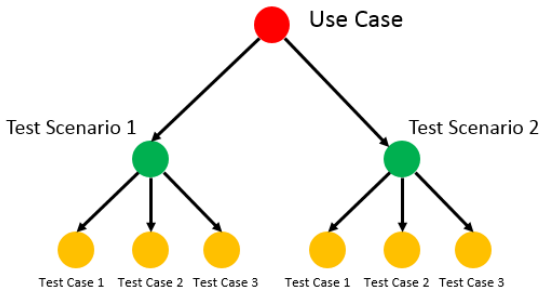

Figure 6.1: Schematic of pyramid relationship between use-case, scenario and test case

These interactions can be captured as use cases which *"describe the system behaviour as a sequence of actions linking the result to a particular actor"* (e.g. driver). Subsequently, scenarios (a specific sequence of a use case) present possible ways in which a system may be used to accomplish a desired function. However, writing scenarios require detailed domain knowledge, which is only found with experts. Moreover, the term "use cases" and "scenarios" have been used with a fuzzy meaning (Campbell, 1992; Cockburn and Fowler, 1998). A use case is a collection of scenarios bound together by a common goal (Cockburn and Fowler, 1998) and implies "the way in which a user uses a system" (Cockburn, 1997).

Scenarios has been suggested to have at least four different meanings: 1) scenarios to illustrate the system 2) scenarios for evaluation 3) scenarios for design 4) scenarios to test theories (Campbell, 1992). It is worth elucidating that a scenario that is good for illustrating a system (demonstration) may not be good for evaluating the basic functions (requirement

based testing), as the former only uses a limited number of examples. Similarly, scenarios to test theories establish the strengths and more importantly the weaknesses of a design. Therefore, they go beyond the traditional requirement based testing.

6.1.1.3. Discussion: Identifying the research question and research objective

It has been suggested that to prove the safety of a vehicle with an automated system, it is not about the number of miles that are driven by the vehicle but about what the vehicle experiences in those miles (Wachenfeld and Winner, 2017a). However, defining and identifying "relevant" miles or "good" miles remains a challenge for the research community and the industry.

One of the reasons behind suggesting the requirement of 11 billion miles to prove the safety of the system is that it would cover all possible *"black swan"* and known unknown scenarios. However, *"black swan"* scenarios by its very definition may or may not be covered even with 11 billion miles. Therefore, rather than focussing on the number of miles, the focus of research should be on the identification of the *"black swan"* and *"known unknown"* scenarios.

While requirements based testing captures the *"known knowns"* efficiently, the inability to ensure its completeness leads to the occurrence of *"unknown knowns"*, *"known unknowns"* and the *"black swan"* scenarios. The three later together could possibly be combined to form "good" scenarios and to define such scenarios should be the goal for testers.

Considering in the context of the literature discussed on trust, the experience of an unknown failure by the driver has a potential of causing a large decrease in trust leading to disuse of the system. Additionally, as discussed in chapter three, recovery of trust (once lost) is a much slower process as compared to the initial development of trust. Therefore, the objective for the design of automated systems should be to inform the driver about the true capabilities and limitations, i.e., prevent the driver from experiencing *"black swan"* or *"known unknown"* or *"unknown known"* scenarios.

While the aim of testing should be to increase the coverage of *"known known"* scenarios by better specification, one of the challenges of identifying *"black swan"* scenarios is their lack of correlation with time (Wachenfeld and Winner, 2017a). Interestingly, the lack of consensus in defining a *"black swan"* or a *"known unknown"* scenario leads to the presumption that there is a lack of understanding on the definition of a "good" test scenarios. Based on the review of the literature, the author identifies the following research question and research objectives:

Research Question 2

How to create test scenarios to establish the limitations of the automated driving systems?

In order to facilitate the process of answering research question 2, the following research objectives were identified: (in addition to the earlier two research objectives identified in chapter three (literature review on trust in automation) which are associated with research question 1)

Research Objective 3

"To develop an understanding about the characteristics of a test scenario (for ADAS and ADS)"

In order to develop *"good"* test scenarios, the first step requires an understanding of what is meant by *"good"*. Is just defining the *"known known"* or having a structured approach to capture the *"unknown knowns"* is enough? Or is there a need to capture more of the *"black swan"* scenarios within the realms of a *"good"* test scenario description? Additionally, it is important to understand whether the characteristics differ for an automated system as compared to a traditional automotive system e.g. ECU (engine control unit), TCU (transmission control unit) etc.

Research Objective 4

"To develop a methodology for creating test scenarios"

Once the characteristics of a "good" test scenario have been defined for ADAS and ADS, it is important that the methodology to create test scenarios incorporates the ability to generate test scenarios with the identified characteristics.

6.1.1.3.1. Next research step

Existing requirements based testing approach widely used in the industry, only ensures that the system meets its requirements while failing to identify the exceptions explicitly. Some exceptions may be covered sporadically due to the experience of historic failures rather than a scientific approach. Additionally, there is a known challenge to ensure completeness of requirements. Requirements reflect the expert's view of system's functionality and possible usage. The identification of the requirements has a degree of subjectivity associated with this (Flage and Aven, 2015). Different experts with different background knowledge analyse and classify systems differently, leading to inter-rater variation (Ergai et al., 2016).

This is evident in the variation in the classification and identification of scenarios like the *"Black Swan"* scenarios or the *"unknown unknowns"* (scenarios that we don't know that we

don't know) associated with the functionality of the system (Aven, 2013) and the variation in the safety analysis associated with risks associated with those scenarios. While it is important to identify and classify scenarios as a part of testing, it is equally important to do so in a reliable manner. This aspect of testing is discussed in section 6.1.2. Reliability of knowledge becomes an influencing factor on the development of trust when dynamic (real-time) knowledge is provided to the driver. As driver receives real-time knowledge continuously, differing information to same situations lead to the development of distrust causing the driver not to use (disuse) the automated system.

6.1.2. Safety Analysis

6.1.2.1. Reliability and validity of safety analysis

Safety analysis can be divided into two-steps. First step involves the identification of the hazards on which the Hazard Analysis and Risk Assessment (HARA) will be performed. There exist many methods for identifying hazards like System Theoretic Process Analysis (STPA) / Systems Theoretic Accident Model & Processes (STAMP) (Leveson, 2004; Nancy G Leveson, 2011; Nancy G. Leveson, 2011), JANUS (Hoffman et al., 2004), Accimaps (Salmon et al., 2012), Event Tree Analysis (section 7.2.2), HFACS (Baysari et al., 2009; Chen et al., 2013; D. Wiegmann and Shappell, 2001b), Fault-tree analysis (Lee et al., 1985; Reay and Andrews, 2002), bow-tie analysis (Abimbola et al., 2016; Khakzad et al., 2012), FMEA (Stamatis, 2003), etc. Some of these methods were developed for simpler systems and fall short in their ability to meet the requirements for the analysis of modern systems which have multiple interactions between the system and software components and the human operator (Fleming et al., 2013). Another source of identifying hazards is from experience of previous accidents and their accident investigations. However, being retrospective in nature, they cannot be taken as the only source of possible hazards, but should influence future hazard identification process and safety management process (Stoop and Dekker, 2012). While accident investigations provide new knowledge about the possible avenues of system failures, they are never exhaustive. This is evident by the deja-vu experience of similar accidents repeating themselves in a 20-30 year cycle (Le Coze, 2013). Identifying hazards has its challenges and is a research question in its own right. While it is possible to identify hazards based on the *"known knowns"* and accommodate for the *"known unknowns"*, it is extremely difficulty to foresee the unknown knowns and even more so for the *"unknown unknowns"* which form the *"Black Swan"* category for hazards (Aven, 2013). Previous accidents, however, provide an insight to the occurrence of *"Black Swan"* type of accidents by increasing experts' knowledge of possible factors for risk analysis (Khakzad et

al., 2014). Answering RQ2, the author discusses the procedure for identification of hazards and hazardous scenarios in chapter 7.

The second step of the safety analysis process involves the analysis of the hazard and the corresponding risk assessment for the hazard. Risk in general has been suggested to be a construct and not an attribute of the system (Goerlandt and Montewka, 2015), due to the subjective nature of risk (Aven, 2010a; Tchiehe and Gauthier, 2017). However, in the automotive domain, a decomposition of risk provides a different insight. An Automotive Safety Integrity Level (ASIL) rating in automotive HARA comprises of a severity, exposure and a controllability rating. Controllability and Severity of any system are a system attribute. However, exposure for a system remains a construct and is open to subjective variation as it is influenced by the expert's knowledge which governs the probability rating (Aven, 2010b; Aven and Reniers, 2013). Automotive HARA and ASIL will be discussed in detail in section 6.1.2.2 and chapter eight.

While HARA governs the risk management, i.e., the mitigation steps and the rigour required in the application of the steps; it is plagued by some fundamental challenges of its validity and reliability (Aven and Zio, 2014). One of the fundamental issues with risk assessment is the biases or assumptions made by stakeholders performing the assessment due to subjective interpretation of the underlying process or lack of knowledge of the underlying uncertainties or lack of knowledge of the system safety. Lack of knowledge or improper knowledge about the system may lead to either ignoring possible risk (which may lead to false negatives) or their exaggeration (which may lead to false positives). This introduces uncertainty in the risk analysis which is not taken into consideration while making decisions (Goerlandt and Reniers, 2016). Additionally, the knowledge of the hazards and possible failures helps guide the design process of the systems by providing the ability to make informed design decisions in the design phase of the product (Björnsson, 2017; Villa et al., 2016).

Reliability refers to the *"extent to which a framework, experiment, test, or measuring instrument yields the same results over repeated trials"* (Carmines and Zeller, 1979). In a review of Quantitative Risk Analysis (QRA) method applications, (Goerlandt et al., 2016) found that significant differences existed in the results of QRA conducted by different teams/groups of experts. While mandating a specific QRA method could reduce variation (Van Xanten et al., 2013), they argued that this would not ascertain the accuracy of the results, but make results converge and more comparable.

For HARA to be scientific, it needs to be reliable (Hansson and Aven, 2014). In this thesis, the author adopts the "reliability" definition and types of reliability as defined by Aven and Heide (2009) (pg. 1863):

- *"The degree to which the risk analysis methods produce the same results at reruns of these methods (R1)*
- *The degree to which the risk analysis produces identical results when conducted by different analysis teams, but using the same methods and data (R2)*
- *The degree to which the risk analysis produces identical results when conducted by different analysis team with the same analysis scope and objectives, but no restrictions on methods and data (R3)"*

6.1.2.2. Automotive Functional Safety

In the automotive domain, the ISO 26262 standard (automotive functional safety international standard) lacks a quantified and a robust process for automotive certification (Yu et al., 2016). Even the latest version (2018) of the standard suffers from the same issue. The standard refers to ASIL as a metric for hazard analysis which is influence by Severity (S), Exposure (E) and Controllability (C) rating. However, the methodology for determining these parameters and their quantification is not mentioned. Instead a set of sample tables have been provided (Ellims and Monkhouse, 2012). SAE J2980 provides some guidance to certain degree of objectivity to automotive HARA. But it too falls short in defining various aspects influencing severity, exposure and controllability rating (SAE International, 2015). SAE J2980 provides one table to parametrise severity using speed and collision type as parameters. It doesn't provide any guidance for controllability and exposure ratings. Even for severity, the parameters used are not exhaustive enough.

Thus, there is a need for creating a method for extracting patterns and creating templates for safety case development which would influence the HARA (Kelly, 2004). While ISO 26262 (2011) - Part 3 (ISO, 2011b) comprehensively describes the hazard analysis and identification of hazards using various methods like HAZOP (Cagno et al., 1960), FMEA etc.; it falls short of identifying an objective rating methodology for the hazardous events identified. This leaves the rating to the skills and the mental model of the domain technical experts performing the rating task. An expert's mental model is created and influenced by their own knowledge, experience and environment, leading them to base their risk analysis on some underlying assumptions (Rosqvist, 2010). Any risk rating given by an expert is dependent on the expert's interpretation of the background knowledge (based on their mental model) related to the hazard. This background knowledge may be incomplete in three specific areas: structure of the hazard, parameters responsible for the hazard and probabilities for the parameters (Aven and Heide, 2009). Thus, the mental model formed by the expert is a limited representation of the real world. In addition, the dominance of various

factors influencing expert's mental model differ at different points in time for the same expert, leading to a varying decision making analysis. Thus, the following two types of variations exist in industry when hazard analysis and risk assessment is performed:

- Inter-rateability variation: due to different mental models between different experts or different groups of experts
- Intra-rateability variation: due variation in mental models of the same expert or same group of experts at different points in time

In a study to evaluate the reliability of the Human Factors Analysis and Classification System (HFACS) (Shappell et al., 2007; D. A. Wiegmann and Shappell, 2001), which is a retrospective accident analysis framework, it was found that while training of experts improved reliability of the analysis, the results demonstrated significant inter- and intra-rater variation (Ergai et al., 2016). Even classification of a hazardous event as a *"black swan"* is of subjective nature and is prone to inter-rater variations. It is also influence by knowledge or beliefs of the experts which is based on their individual mental models (Aven, 2015; Flage and Aven, 2015).

In order to overcome this challenge, an approach would be to increase focus on the knowledge aspect of HARA by having two teams independently performing the HARA. The role of the second team being to check the bias and the assumptions made by the first team (Veland and Aven, 2015). While such an approach has its merits, it is not practical to adopt this approach in the automotive industry due to the time and human resource required for the approach. The automotive industry is overwhelmed by time and cost constraints to meet production deadlines, therefore a novel approach is required for addressing the reliability issues of the automotive HARA process, while meeting constraints of the automotive industry.

6.1.2.3. Discussion: Identifying Research Question and research objectives

While existing literature acknowledges the reliability issues, a solution to tackle the inter- and intra-rater variation still evades the research community. The work presented in this thesis (chapter eight) focusses on increasing reliability of the automotive HARA process by objectivising the severity, exposure and controllability ratings by introducing a rule-set for both the ratings. This work is one of the first steps towards achieving reliable ratings through an objective decision making process for HARA in order to provide reliable dynamic knowledge to the drivers to calibrate their trust appropriately. The (third) research question (addressing the reliability issues of the automotive HARA) focussed in this thesis is:

Research Question 3

"How to improve the inter- and intra-rater-reliability of the automotive HARA process?"

In order to facilitate the process of answering research question 4, the following research objectives were identified (in addition to the four earlier research objectives identified in chapter 3 (RO 1 and RO 2) and section 6.1.1.3 (RO 3 and RO 4)):

Research Objective five

"To develop a rule-set for conducting automotive HARA"

In order to improve the reliability of the automotive HARA, the author proposes to objectify the process to remove the variation caused by subjective interpretation of the process. Creation of the rule-set involved parametrization of the severity, exposure and controllability ratings.

Research Objective six

"To determine the ability of the developed rule-set for HARA in improving the reliability of the automotive HARA"

After developing the rule-set for severity, exposure and controllability, the rule-set needs to be tested to evaluate its ability to improve the reliability of the automotive HARA process. Inter-rater reliability is discussed within the scope of this thesis.

6.2. Summary

The two driving simulator studies in chapter five demonstrated that by introducing knowledge (static and dynamic) it is possible to calibrate drivers' trust to prevent disuse and misuse of automated systems. In order to create this knowledge in a reliable manner, automated systems need to be tested in a manner that generates reliable results. Testing of automated systems comprises of: 1) using test method(s) 2) identifying scenarios and 3) safety analysis (classifying the risk in the scenarios). The latter two have been discussed in this chapter.

With the advent of ADAS and ADS in vehicles, the complexity of vehicles has increased tremendously. The number of lines of code in a modern luxury car is over 100 million as compared to a 6.5 million in a Boeing 787 airplane (Charette, 2009; Strandberg et al., 2018). This increased complexity has led to challenges in testing of the safety critical automotive systems (i.e., ADAS and ADS). Coupled with relatively short development cycles in the

automotive domain as compared to the aviation or rail, there is an urgent need to develop innovative methods to test ADASs and ADSs.

Discussing test scenarios, it has been suggested that automated driving systems need to be driven for 11 billion miles in order to statistically prove that they are safer than human driving vehicles. While this is an infeasible proposition, the industry and the research community do not yet have a solution for identifying test scenarios. One school of thought suggests that even after 11 billions miles one cannot guarantee the identification of all possible *"black swan"* and known unknown scenarios.

Automotive HARA which classifies risk in various hazardous scenarios suffers from reliability issues. This is due to the differing mental models of the safety experts, which is based on their assumptions, experience, culture and many other factors. Thus, the author has identified an unanswered question in current literature on test scenario creation and the research gap concerning the reliability of the current automotive risk analysis methods.

6.2.1. Research Questions and Research Objectives

Based on the review of the literature on test scenarios (in section 6.1.1) and reliability of HARA (in section 6.1.2), the following research questions (RQ) and their corresponding research objectives (RO) have been identified:

RQ 2. How to create test scenarios to establish the limitations of automated driving systems?

- *RO 3. To develop an understanding about the characteristics of a test scenario (especially for ADAS and ADS)*
- *RO 4. To develop a methodology for creating test scenarios (based on the identified characteristics)*

RQ 3. How to improve the inter- and intra-rater-reliability of the automotive HARA process?

- *RO 5. To develop a rule-set for conducting automotive HARA*
- *RO 6. To determine the ability of the developed rule-set for HARA in improving the reliability of the automotive HARA*

In the next chapters (chapter seven and chapter eight), the author discusses the research methods and outcomes corresponding to the research questions and the research objectives identified in this chapter.

HOW TO CREATE THE CONTENT FOR INFORMED SAFETY?[4]

Chapter 7

The recent crashes involving Uber's ADS equipped vehicle (NTSB, 2018) and Tesla's Autopilot equipped vehicle (NHTSA, 2017b), demonstrate the need for a framework to test for the *"black swan"* scenarios and the *"known unknown"* scenarios. From a statistical point of view, in order to prove with 95% confidence that autonomous vehicles (AVs) are even 20% safer than human driven vehicles, they need to be driven for over 11 billion miles (Kalra and Paddock, 2016a). While this seems unfeasible to achieve, it is important to understand what kind of miles should constitute these 11 billion miles. It is possible to make safety claim by driving 11 billion miles on a straight road on a sunny day in the middle of the desert. However, such testing may not be sufficient or even relevant for ADAS and ADS to be deployed in environments where Vulnerable Road Users (VRUs) are present, or where it rains/snows heavily. Therefore, the number of miles driven by the ADS or ADAS are not as important, as is what the ADAS or ADS experience in those miles. Additionally, it is important to test the ADAS or ADS in the operational design domain it has been designed for (e.g. a low-speed automated shuttle designed for urban environment should be tested in urban settings and not in the middle of a desert).

In this chapter, research question 2 (RQ2: How to create test scenarios to establish the limitations of automated driving systems) is answered by meeting research objectives three and four (RO3 and RO4).

[4] Contents of this chapter have been published in the following publication:

Khastgir, S., Birrell, S., Dhadyalla, G. & Jennings, P. The Science of Testing: An Automotive Perspective. in SAE Technical Paper: 2018-01-1070 (2018). doi:10.4271/2018-01-1070

7.1. Understanding characteristics of test scenarios (for ADAS and ADS): A semi-structured interview study

Answering RQ2, in order to understand the test scenario creation approach being undertaken by the automotive industry towards ADAS and ADS to uncover the *"unknown unknown"*, *"unknown known"* and *"known unknown"* scenarios, a semi-structured interview study involving verification and validation experts in the automotive domain was conducted. Semi-structured interviews were conducted to understand the existing knowledge base for test scenario generation process in the automotive industry and their understanding and expectations from an ideal test scenario for ADAS and ADS. While it is impractical to test for 11 billion miles, it is important to identify an approach to create test scenarios for which the system needs to be tested

Semi-structured interviews provide the ability to have richer data from the interviewees (as compared to surveys and questionnaires) (Louise Barriball and While, 1994). Additionally, they allow the flexibility to examine topics in different degrees of depth (as per interviewees' interest and background) (Robson and McCartan, 2016). The interviews were transcribed and the text was sanitized to remove any proprietary mentions. A coding analysis was performed on the sanitized text and themes and categories were identified from the various interview answers.

In order to prevent any bias, the interviewees were allowed to talk freely while answering the questions and were not prompted for any answers. Participant interviews were transcribed into text and were later coded to perform thematic analysis. Key themes were identified from the transcribed text.

7.1.1. Method

Ethical approval for the study was secured from the University of Warwick's Biomedical & Scientific Research Ethics Committee (BSREC) (BSREC Application ID: REGO-2016-1824). All interview transcripts were anonymized and stored in a secure location and University of Warwick's data handling procedures were followed.

7.1.1.1. Participants

Twenty industry experts with each participant having over 10 years' experience in the field of testing and development of systems in the automotive industry were recruited for this study. Participants were selected from a diverse demography cutting across the automotive

supply chain. Eleven participants represented OEMs (Original Equipment Manufacturers), eight participants represented Tier 1/2 suppliers and the remaining participant represented academic /research organizations /start-ups working in the area of automated driving. To ensure independence of the interviewees, participants were recruited from different countries including the UK, Germany, India, Sweden and USA. The interviews lasted between 28.63 min and 103.15 min (average interview length: 48.25 min). Interviewees were also assured that any of the responses will not be identifiable to them as the transcripts would be anonymized before they were analysed.

7.1.1.2. Interview questions design

The interview was structured with six guiding questions, which were divided into two themes: 1) test methods (three questions) 2) test scenarios (three questions). Each guiding question had a set of follow-up questions, which were asked depending on the content of the answers. The six guiding questions were the following:

Test Methods

1. What test methods do you use for testing of automotive systems?
2. What are the challenges for each test method that you have faced?
3. What metric do you use to measure sign-off criteria for testing automated systems?

Test Scenarios

4. How do you ensure robust testing of automated automotive systems to various driving conditions?
5. How do you develop test scenarios for testing automated systems?
6. What criteria do you think makes a good quality test scenario?

The set of follow-up (prompting) questions are described in Table 7.1. The follow-up questions were used to aid participants thought process and were designed to be minimally prescriptive to avoid biasing the answers.

Table 7.1: Interview follow-up questions

Guiding question #	Follow-up Question(s)
1	Reason for selecting a test method? What tools do you use as a part of your test setup?
2	What is your biggest challenge?
3	How was the metric developed? Is it a standard metric? (Company internal or industry standard)
4	What test scenarios do you use while doing real world / virtual testing?
5	What aspects are critical while developing a test scenario for autonomous system?
6	How did you develop those (for good quality test scenario) criteria?

7.1.2. Data Analysis

As this study employed a semi-structured interview format, the analysis of the data was mostly qualitative. In order to structure the data analysis and identify trends in the collected data, a coding strategy was used. A code *"is a word or short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute to a portion of language-based or visual data"* (Saldaña, 2016). By reading through the transcribed interview text, codes were assigned to the text which enabled conversion of the interview text into an easy to understand tabulated format. An example of a code and corresponding text is discussed here. One of the responses to the question on the biggest challenge in testing faced by the interviewee was, *"it is difficult to create the specification to verify against and because of the lack of specification, it is difficult to put a criterion for completeness of testing"*. The corresponding code assigned to the text was *"how to ensure completeness of requirements"* and *"how to judge test completeness"*. However, it is evident that such a coding process is subjective due to the understanding and biases of the coder. In order to overcome this, a two stage coding process was followed which was reviewed by an independent person.

7.1.2.1. First cycle coding

The first coding cycle involved reading through the interview transcript to assign codes. As the data in a semi-structured interview transcript can be varied, different methods of coding such as structural coding, descriptive coding, process and in-vivo coding, were used (Saldaña, 2016).

7.1.2.2. Second cycle coding

Since different coding methods were used in the first coding cycle, some of the codes were similar or split. In order to synthesize the first cycle codes to develop a more cohesive understanding, axial coding was used in the second phase which led to the creation of categories for the first cycle codes. Table 7.2 illustrates the coding process for the answers received to question five.

Table 7.2: Development of codes for question five

Participant answers	1st cycle codes	2nd cycle codes
"what kind of environmental influences could lead to an ill function"	Environmental factors, failures	- Identify failures, system limits, and hazards. - Using **systematic method** to identify failures, hazards
"try to define a test to see the degradation of the performance"	Degraded performance, faults	
"understand situation in which our system will reach any kind of limit "	System limits, degraded performance	
"fidelity of test scenario comes down to the FMEA. Because out of the FMEA there is possibility of a failure you need a control method for"	Identify failures, systematic way, FMEA	
"we would engineer faults into the system…. Blocking the radar. Put radar absorbent material (RAM) for the radar."	Create faults, block sensors	
"it is currently done via FMEA, System FMEA and Hazard Analysis"	Systematic way, FMEA, HARA	- Using **systematic method** to create test scenario library
"have a catalogue of tests"	Test library	
"systematic way (of) what kind of influencer I have into the behaviour of the functionality…"	Factors influencing functionality, systematic method	
"you can think about you have a matrix…then we look at what kind of combinations are possible"	Test library	

7.1.3. Results

While it was found that tools (software platforms) for test execution were not an issue for most organizations, the infrastructure requirement for such test platforms (hardware-in-the-loop setup and instrumented test vehicles for real-world testing) had exponentially increased with ADASs and ADSs as compared to traditional automotive systems. In addition, the large amount of data handling required for sensors used in the ADASs and ADSs was another challenge.

In response to the first question on test methods used for testing, the participant responses could be grouped in two themes. One group of participants commented that they follow the software development V-cycle and implemented model-based design tools using simulation in a major part of their development process. On the contrary the other group was of the opinion that simulation is of limited use for ADASs and ADSs as it is "almost impossible" to model sensors, especially RADAR and LiDAR sensors and they mostly depended on real world testing.

More importantly, the input to the test execution platform (test case vectors) was a common concern acknowledged by all participants. When asked about the biggest challenge faced by the participants while performing testing, two specific themes emerged. While the OEMs credited "test case generation and definition of pass/fail criteria" as their biggest challenge; tier 1/2 suppliers credited "quality of requirements (including completeness and consistency)" as their biggest challenge. This difference can be credited to the culture in the automotive supply chain where the suppliers develop individual systems and the responsibility for integration of these systems lies with the OEMs. However, both the groups failed to mention any solutions to the challenges faced by them during the testing phase; the ability to identify and define the "known unknown" and the "unknown unknown" scenario space.

When asked about the parameters and criteria for *"good"* test scenarios, there seem to be an agreement on the ability to test "known unknown" and "unknown unknown" situations, as a key feature of a "good" test scenario. However, a deeper analysis of the responses revealed two distinct themes. Firstly, creating scenarios from requirements is dependent on the skill and experience of the test specifiers. Secondly, *"good"* scenarios should be able to test safety goals and ways in which the system may fail. This is generally not covered by system requirements. Moreover, the need for a systematic method of identifying the system limits or failure scenarios was highlighted by the participants. Most experts mentioned that Requirements Based Testing (RBT) is insufficient as there is a challenge in ensuring completeness of requirements. RBT captures the typical scenarios as suggested by the requirements and represent the most common real world scenarios. Such testing ensures that the most common bugs are identified (Whittaker, 2000). While approaches to improve requirement based testing have been discussed in literature (Robinson-mallett et al., 2010; Robinson-mallett, 2012), discussion on the ability to increase the *"known known"* by identifying the unknown space is limited. One of the reasons mentioned by experts about RBT was that it is impractical to have a requirements document capturing the multitude of scenarios an automated driving system might encounter, rendering the classical V-cycle for software development obsolete.

In the testing process, it is important to establish when to stop testing and sign-off the system-under-test. When the participants were asked about a metric used to measure the sign-off criteria, surprisingly the answers demonstrated the lack of any standard metric in place. Unfortunately, the sign-off point was dependent on the budget allocated and SOP time. However, all participants acknowledged that this wasn't the ideal situation and needs to change for ADASs and ADSs. However, some participants did provide some insight into an ideal situation and using false positive and false negative rates as metric for sign-off.

When asked about how participants ensured that the ADASs and ADSs were tested robustly, they mentioned using a test catalogue which was developed from experience. However, all participants agreed that for ADASs and ADSs, more real world testing is needed due to the challenges in simulation environment. On the time split between real-world and virtual testing, one of the participants commented: *"95% is real world testing and 5% is simulation. But for me it should 50-50. For the moment the robust model of the simulation is stopping (this to happen)"*.

7.1.4. Discussion

One of the challenges of identifying *"black swan"* scenarios is their lack of correlation with time (Wachenfeld and Winner, 2017b). Based on the analysis of the interviews, to increase the area covered by the "known knowns" in the test scenario space, the author proposes a two-pronged approach to testing of ADASs and ADSs to create test scenarios and test cases (Figure 7.1). The first branch concerns using traditional RBT approach, while the second branch uses a Hazard Based Testing (HBT) approach for creating test scenarios. Traditional RBT method covers only a fraction of the possible test scenario space for the systems (Figure 7.1). The addition of the second testing branch (HBT) improves the coverage of the test scenario space by increasing the "known known" scenario space. However, it does not guarantee full coverage of the test space (Figure 7.1). While RBT checks the working of the system as per expectations (defined requirements), HBT explores how the system may fail by identifying possible failure scenarios.

HBT draws its inspiration from the world of security analysis. In security analysis, the use of misuse cases has been suggested has a way of testing for security concerns (Alexander, 2003). Misuse cases can *"help document negative scenarios"* (Alexander, 2003). The key to the success of HBT is to have a structured, robust and well-documented method of identifying hazards. This was also highlighted in the themes obtained from the analysis of the interview transcripts. The two themes were: *"failure or hazard scenarios"* and *"systematic method (objective) to obtain them"*. On being asked about how to develop test scenarios, one of the participants commented: *"try to define a test to see the degradation of*

the performance", while another participant mentioned: *"what kind of environmental influences could lead to an ill function"*.

In order to identify hazards, various methods like HAZOP , Fault Tree Analysis (FTA) (Lee et al., 1985; Reay and Andrews, 2002), FMEA (Stamatis, 2003), Event Tree Analysis (ETA), JANUS (Hoffman et al., 2004), Accimaps (Salmon et al., 2012), HFACS (Baysari et al., 2009; Chen et al., 2013; D. Wiegmann and Shappell, 2001b), bow-tie analysis (Abimbola et al., 2016), System Theoretic Process Analysis (STPA) / Systems Theoretic Accident Model & Processes (STAMP) (Leveson, 2004, 2012) etc. have been used in the industry and research community. Some of these methods were developed for simple systems and fall short in analysis when dealing with the complex nature of interactions between system, human operator and the software in modern ADAS and ADS (Fleming et al., 2013). As an example, in the case of an Adaptive Cruise Control (ACC) system, rather than testing the functional requirements, more emphasis needs to be laid on identifying the hazards associated with the usage of an ACC system. An analysis of the ACC system using any of the earlier said methods to identify hazards would lead to one of the potential hazards as *"unintended braking"*. Some of the hazard identification methods developed specifically for ADAS and ADS (e.g. HFACS, JANUS, STPA) further analyse the system interactions to identify that one of the potential cause of an *"unintended braking (hazard)"* could be the "vehicle manoeuvring through a steep bend" causing the radar system to believe that there is an obstacle in front. Therefore, an HBT approach would identify such situations that would have been missed in a traditional RBT approach.

In order to identify the safety goals and the hazards, a Hazard Analysis and Risk Assessment (HARA) process needs to be conducted. The automotive HARA has its own issues like subjective variation due to skill and experience of the testers and completeness of the HARA, some of these issues have been answered in the literature (discussed in chapter six and eight). Once the systems have been tested, their capability and safe performance can be correctly established and can form part of the knowledge to be imparted to the drivers, in real time or before they start their usage, establishing their "informed safety" level to improve trust in ADAS and ADS (discussed in chapter 5). However, in order to create the "informed safety" level, a more systematic and structured process needs to be adopted to testing. As one of the interviewees mentioned, *"Testing is a science"*.

Due to practical reasons, this study had a limited sample size for the interview pool. While a large number of practitioners are involved in the field of verification and validation, it would be a major challenge to interview a representative sample size. However, due to the expertise of the interviewees, current findings (Hazard Based Testing) provided an important insight

for this research and testing of ADAS and ADS. The proposed Hazard Based Testing approach comprises of three steps: 1) Identification of Hazards 2) Creating test scenarios for the identified hazards 3) Risk classification of the identified hazards. In this chapter, first two steps of HBT are discussed. The third step of HBT will be discussed in chapter eight.

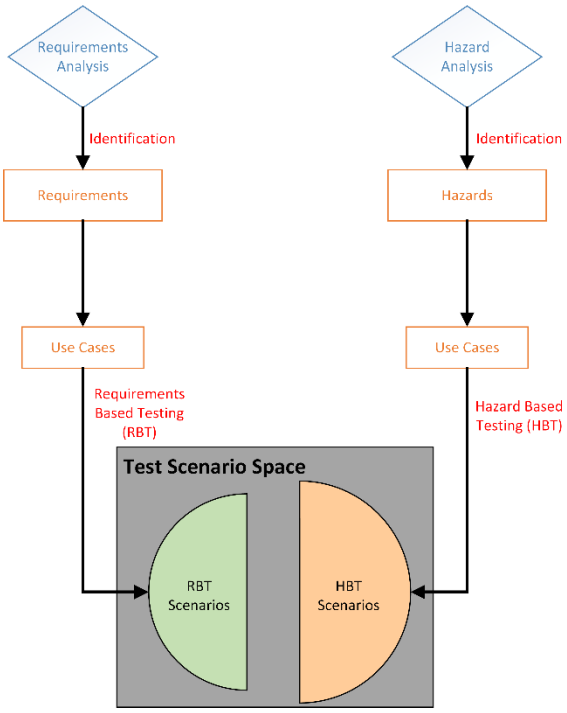

Figure 7.1: Proposed testing approach for test-scenario generation

7.2. Identifying hazards

Having determined that there is a need to identify hazards in order to conduct Hazard Based Testing (HBT), the author will now discuss various methods used to identify hazards and their causal scenarios. Hazard analysis methods help in identifying future causes of accidents for a system under analysis. While there are various hazard analysis methods, the author will discuss some of the more commonly used methods in the next sections. These methods include Failure modes and effects analysis (FMEA), Fault Tree Analysis (FTA), Event Tree Analysis (ETA), HAZOP and Systems Theoretic Process Analysis (STPA). This section provides a brief overview on each of these methods. While discussing these methods, the strengths and limitations of each of the methods will be highlighted, ultimately concluding with why STPA was chosen for the hazard identification process for the purpose of this thesis.

7.2.1. Fault Tree Analysis (FTA)

Fault Tree Analysis (FTA) is a versatile method which was conceived in 1961 at Bell Laboratories by H.A. Watson to analyse the Minuteman missile launch control system, under a US Air Force contract. It was created as at that time there was no method to analyse electro-mechanical failures leading to hazardous situations e.g. accidental missile fire. Soon after its creation, it gained widespread adoption with Boeing adopting FTA as a part of its aircraft design process. As of today, FTA is one of the most popular methods used for hazard identification and hazard analysis.

FTA is a deductive, top-down approach based on analysing chain of events, especially combination of events which could potentially cause a hazard, e.g. accidental missile launch. Based on reliability theory, Boolean algebra and probability theory, FTA translates system's failure behaviour into a visual diagram. As the name suggests, FTA creates a tree-structure for events and failures (Figure 7.2). At top of the tree is an undesirable event or a fault condition identified by the person performing the FTA. Subsequently analysis progresses in a top-down manner applying Boolean logic (e.g. AND, OR etc.) to component failures and events.

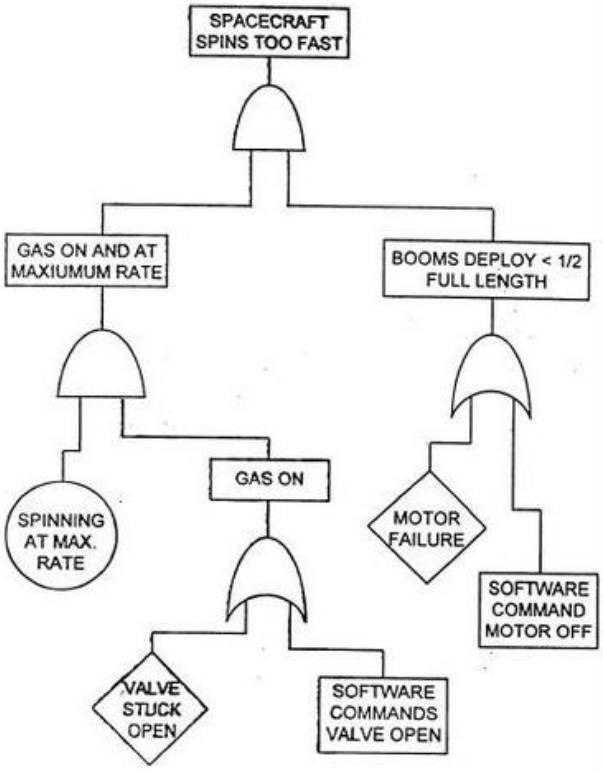

Figure 7.2: Example of a High Level Fault Tree
(Image reference: https://commons.wikimedia.org/wiki/File:Example_of_High_Level_Fault_Tree.jpg)

As FTA gained popularity with its successful application in the missile program, FTA practitioners increased and further developed FTA with more formal rules and methodology (Kaiser et al., 2003; Vesely and Roberts, 1981). FTA has been successfully applied to systems with well-defined failure modes which has meant that accidents caused by component failures have mostly been mitigated. One of the variations of FTA – Quantitative FTA, includes assigning probabilities to the occurrence of the failures and the events. Such probabilities are mostly done based on expert judgements with little or no objective data to back it (Aven, 2010b; Aven and Reniers, 2013; Mosleh et al., 1988; Rae and Alexander, 2017).

As systems become complex with more software component and increasing number of electro-mechanical components, new types of accidents have emerged which are caused due to: 1) component interactions 2) software errors/bugs. While the former is driven by incomplete requirements and design errors, the latter can have infinite possibilities. Software bugs are interesting as software in itself never fails (Nancy G. Leveson, 2011). It does exactly as it has been implemented. It is the implementation that be incorrect or the requirements could be incorrect based on which the implementation is made. Both these situations are harder to illustrate in a fault tree as software failures can have infinite possibilities unlike physical components which have well defined failure modes.

Additionally, for ADASs and ADSs, the human driver/passenger is in the loop and needs to be accounted as a possible cause of failure. Incorporating the complex nature of human behaviour and human-automation interaction in the FTA poses another limitation as human behaviour is adaptive and at times unpredictable. While the success of FTA concept depends on the knowledge of failure modes and effects of the systems, the changing nature of systems has meant that this knowledge is limited. As mentioned earlier, at the top of a FTA tree, is an undesirable event which needs to be identified using some other method. The analysis to identify lower level failures is generally a function of the experts' (performing the FTA) expertise and experience. Due to the lack of any guidance from the FTA process on the system model, the analysis will be done based on the expert's mental model leading to not only variation across individuals but also to missing failure identification.

7.2.2. Event Tree Analysis (ETA)

Unlike FTA, Event Tree Analysis (ETA) method is a bottom-up approach which was first conceived during the study of WASH-1400 nuclear power plant in 1974. As FTA as a method was already in existence at that time, initially FTA was used for this purpose. However, performing FTA led to the creation of large fault trees and the team wanted an alternative, thus developing ETA. Like FTA, ETA is also based on analysing a chain of

events using Boolean logic and starts with an undesirable event or a hazardous situation. An ETA tree starts with the identification of an undesirable (failure) event and its subsequent events are evaluated in a binary manner (success or failure) (Figure 7.3). Like FTA, ETA can also be done in a quantitative manner. FTA is used to calculate the probabilities of failure and the mathematical formula ($1 = p(failure) + p(success)$) is used to calculate probabilities of success.

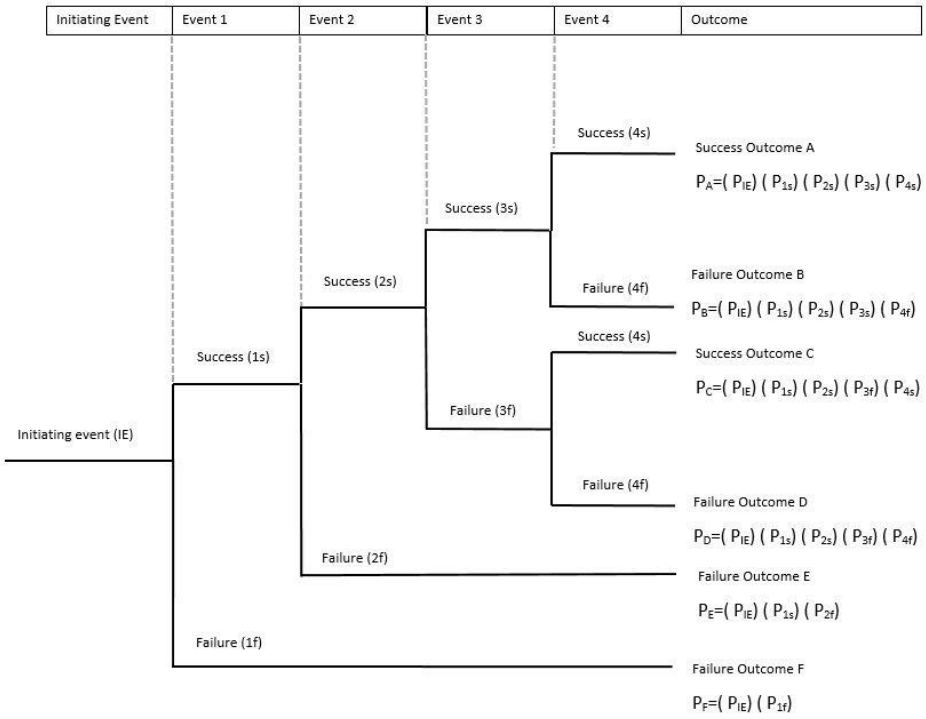

Figure 7.3: Event Tree Diagram
(Image courtesy: Wikimedia Commons)

In an ETA method, it is assumed that the individual subsequent events or barriers to stop propagation of an initial failure, occur independent of each other. This is fundamental to the final outcome probability calculation as it can be calculated by multiplying each individual event probability. However, there have been several examples of catastrophic accidents which have shown multiple failures/events been caused by the same factors (e.g. Three Mile Island accident (Perrow, 1981), Fukushima accident (Labib, 2015)). While ETA also starts with an undesirable (failure) event, it doesn't provide a method to identify the undesirable events in a systematic manner. Thus, another method needs to be used to identify the undesirable event. Like FTA, ETA method, fails to capture issues due to design flaws or incorrect/missing requirements. This has been observed as one of the main reasons for accidents in ADAS and ADS, where assuring completeness of requirements remains a challenge. ETA method aims at mitigating the percolation of the initial failure event down

the tree by evaluating subsequent events or barriers. However, it doesn't focus on any preventive action to block the initiating failure event from occurring in the first place. While human-behaviour can be evaluated within an ETA diagram (in terms of an event or a barrier), it doesn't provide rich analysis due to the binary nature (success or failure) of the event tree analysis. When it comes to human-automation interaction, analysing what caused the human to make that decision can be more important than if the decision was made or not. This granular level of analysis is not considered in ETA.

As mentioned earlier, some industries use FTA in conjunction with ETA. While this provides some benefits in overcoming some of the limitations of either of the methods, it still fails to overcome some of the critical limitations which are common to both (e.g. software bugs/errors, human error).

7.2.3. FMEA and FMECA

In contrast to FTA, Failure Mode and Effect Analysis (FMEA) is an inductive, bottom-up method based on analysing the effect of failures in a system/component in a systematic manner. FMEA was developed to systematically analyse component failures and its impact of system operations and is based on the assumption that accidents occur as a chain of events. An FMEA requires the person performing it to identify all possible failure modes of the components and its potential causes and then evaluate its effect on the system performance. Many a times, FMEA is supplemented by conducting a criticality analysis (CA), which is called as Failure Mode and Effects Criticality Analysis (FMECA). Often FMEA/FMECA is supported by a quantitative analysis where analysts add probabilities for occurrence of the failures into the FMEA/FMECA analysis. Initially introduced in the 1949 in an US Armed Forces Military Procedures document MIL-P-1629 (now MIL-STD-1629A) (Department of Defence - US Govt., 1980), the method has been used extensively by NASA in their space programs including Apollo, Voyager, Viking and Galileo (Kelm, 2010). By the early 1970, FMEA gained traction in the automotive industry with Ford Motor Company introducing its use in its design processes, post the Pinto car recall. More recently, the method has been used widely in other industries like healthcare, food and process industries (Duckworth and Moore, 2010; Faiella et al., 2018; Pawlicki et al., 2016). With different applications, several types of FMEA have been used. These include: Functional FMEA, Design FMEA (DFMEA) and Process FMEA (PFMEA).

In FMEA, the first step involves identifying various components of the system and their corresponding failure modes, where a failure mode is defined as the *"manner in which an element or an item fails to provide the intended behaviour"* (ISO, 2018b). Next, each failure mode is analysed to for its potential cause and effect. When FMECA process is followed, in

addition to the cause and effect identification, a severity and probability of occurrence is also identified for each failure mode. Figure 7.4 illustrates an extract from an example FMEA worksheet of a missile battery, which has been adapted from (Ericson II, 2005).

Failure Mode and Effects Analysis							
System: Missile		Subsystem: Missile Battery				Mode/Phase: Operation	
Component	Failure Mode	Failure rate	Causal Factors	Immediate Effect	System Effect	Severity	Probability
Electrolyte	Leaks out of case	4.1×10^{-6}	Case defect; pinholes	Electrolyte leakage	No power output	Critical	Remote

Figure 7.4: FMEA Worksheet adapted from (Ericson II, 2005)

Like other methods, FMEA/FMECA also has its limitations. As FMEA is a bottom-up approach, it identifies low-level failures. FMEA/FMECA focusses on single failure modes and doesn't capture the effect of combination of failure modes or system interactions. Additionally, as the beginning of the analysis is with the identification of a failure mode, the method doesn't identify hazards that are not related to the failure modes. It is important to note that not all hazardous scenarios are associated to failures and not all scenarios triggered by a failure are hazardous. The resulting scenarios from an FMEA can be both hazardous and non-hazardous. However, the method by its definition treats both types of scenarios equally, resulting in loss of valuable time in analysing non-safety critical scenarios. Similar to FTA, failure scenarios don't capture hazardous scenarios due to missing or incorrect requirements or due to human error. With increasing complexity in automotive systems with ADAS and ADS introduction, such failures are more common and a hazard identification method is needed that can identify such failures.

Human behaviour is adaptive in nature and driven by the goals for a task. Considering human errors as binary failures underplays many important factors that may lead to a hazard. These include: correct human behaviour as per procedural requirements but the procedural requirements themselves are unsafe; incorrect implementation of procedures due to incorrect understanding of them; human mental model conflicts with the defined procedures.

The methods discussed until now assume an independence between failure events. While this assumption makes it easier to assign probabilities to the failure events, the assumption in itself is ill-informed and incorrect. The in-famous Challenge shuttle disaster report found out that the assumption of the failure modes of the primary and the back-up O-rings are

independent was incorrect when put under low temperature and mechanical pressures. Under these conditions, both the O-rings failed simultaneously (Presidential Commision, 1986).

7.2.4. Hazard and Operability Analysis (HAZOP)

Along with FTA and FMEA, Hazard and Operability study (HAZOP) is one of the most popular hazard identification methods used currently in the industry. HAZOP was developed in early 1960s by the chemical processing industry, but now has been adopted in various other industries like mining, nuclear, robotics and automotive.

HAZOP is an organised and a methodical process used to review the design to identify risks that may not have been found by other methods. It is important to note that the full design is needed in order to perform the HAZOP. HAZOP analysis uses guide-words (Table 7.3) and system diagrams for identifying hazards and is carried out by an experienced group of multidisciplinary experts. A guide word is defined as *"word or phrase which expresses and defines a specific type of deviation from a property's design intent"* (IEC, 2016). HAZOP analysis involves comparing the guidewords to a list of system specific parameters (or properties) (e.g. flow, temperature, pressure, speed etc.). The aim of the process is to find possible deviations which may result in a hazard as a result of the discussion between the analysis team. A deviation is formed when a guideword is applied to a parameter. Thus, system parameters (or properties) and guidewords are key to the success of the process. Once the hazards are identified from the deviations, quantitative analysis may be applied for risk assessment (similar to FMEA/FMECA).

Table 7.3: Some Example HAZOP Guide Words (adapted from (IEC, 2016))

Guide Word	Meaning
No	Complete negation of the design intent
More	Quantitative increase
Less	Quantitative decrease
As well as	Qualitative modification/increase
Part of	Qualitative modification/decrease
Reverse	Logical opposite of the design intent
Other than	Complete substitution
Early	Relative to the clock time
Late	Relative to the clock time
Before	Relating to order or sequence
After	Relating to order or sequence

The simplicity of the HAZOP method is one of its biggest advantages. It is easy to learn but the expertise of the facilitator is crucial for good results from the analysis. However, the concept of using deviations to identify hazards enables the method to identify a wider range of hazards (in addition to electromechanical failures). In order identify system parameters or properties and appropriate guidewords a detailed system design is necessary. This means that the HAZOP process will occur late in the development cycle. Thus major changes either will not be done or may be expensive to implement. HAZOP method does a better job of capturing human-errors as compared to earlier methods discussed in this section. However, it fails to capture some of non-trivial errors and the context associated with them. Software HAZOP is a promising technique, but has been criticised for its incompleteness. This stems from the fact that software HAZOP requires the analyst to have a complete understanding of the software interactions and its effects on the wider system, which can be challenging. HAZOP was initially conceived for simple systems like flow-pipes in a process industry which are vastly different in complexities when compared to today's complex control systems in ADAS and ADS. Additionally, as the HAZOP analysis is driven by guidewords, there always exists a chance for some hazards to be missed as they were not associated with the identified guidewords. Lastly, among all the methods discussed until now, HAZOP analysis method takes a much longer time and can thus be expensive (in terms of time and resources).

7.2.5. Systems Theoretic Process Analysis (STPA)

Systems Theoretic Process Analysis or STPA is a hazardous event identification method which is grounded with systems theory and control theory (Leveson, 2004, 2012). It is designed to analyse safety in a socio-technical system with diverse interacting elements (Leveson, 2012). With foundations in systems based approach, STPA identifies a broader range of hazards which may occur due to a variety of reasons including component failures, component interactions, human-error, human-automation interaction, software issues, incorrect requirements and even socio-technical and organisational factors. STPA process is a four step process (Figure 7.5). One of the key benefits of the STPA allows the person performing the analysis to identify hazards and their corresponding requirements, that if implemented would prevent the hazard from occurring. Therefore, it supports to create the preventive action for a hazard and not just its downstream mitigation (as in other methods).

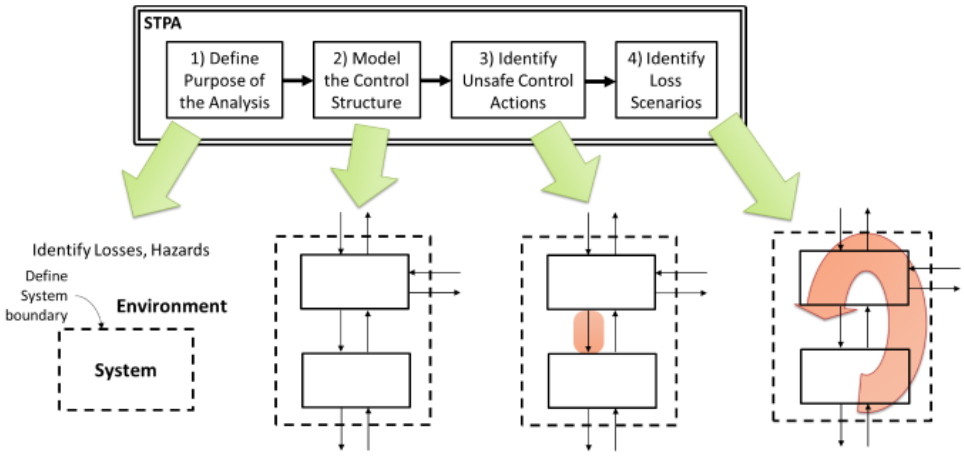

Figure 7.5: Overview of the STPA method (Leveson & Thomas, 2018)

7.2.5.1. STPA Step 1: Defining the purpose of analysis

Step one involves defining the purpose of the analysis. This step involves defining the system (at a higher level) that is to be analysed. It also involves identifying high level "losses" or accidents for the system, which need to be avoided along with potential hazardous states. These losses may include loss of human life or to loss of quality of service (longer trip journey or incomplete trip journey etc.) or may include security, privacy and performance concerns. As an example, for the Low-Speed Automated Driving (LSAD) system that has been analysed as part of this thesis (section 7.4), the following are the high level losses:

1. Crash with a static object or a dynamic actor (L1)
2. Not completing the journey with passenger and cargo (L2)
3. Time of journey being too long, i.e., service target not met (L3)

For ADASs and ADSs, step 1 also includes defining the Operational Design Domain (ODD) of the system.

7.2.5.2. STPA Step 2: Creating system control structure

In step 2 of the STPA method, a hierarchical control structure model of the system is created, capturing the functional interactions by creating a set of nested feedback control loops between the sub-systems. At the beginning, the control structure can be simplistic in nature with basic control actions and feedback loops. Figure 7.6 depicts a control diagram of a simple interlock system. A control diagram generally consists of a controller, actuator,

sensor, controlled process, other controllers, control action, feedback and data inputs. In a typical control structure, controller provides control actions to the actuators and receives feedback from the sensors. The control structure is refined with various iterations as the system requirements are defined and detail is added about the system. It is important to note that the control structure is a not an implementation specification or a detailed software architecture. The control structure is abstracted to a level of capturing functional interactions.

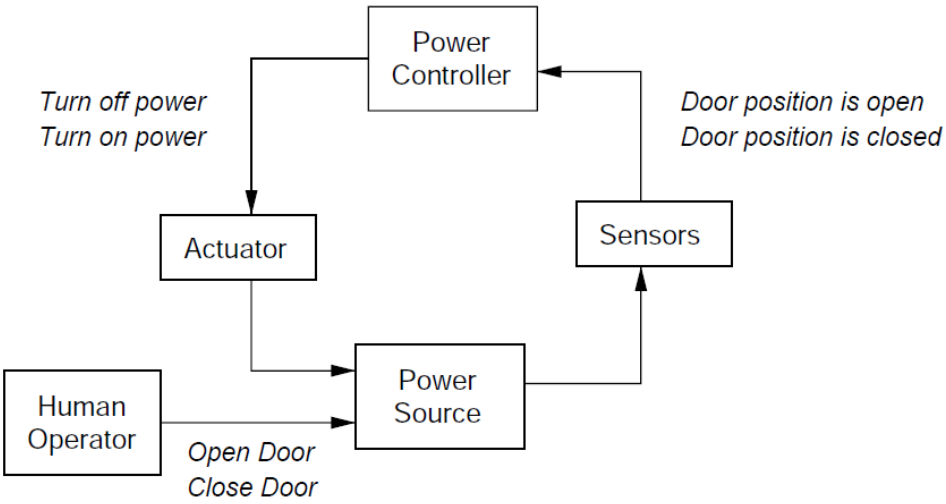

Figure 7.6: Control diagram of a simple interlock system (Leveson, 2012)

The beauty of STPA method lies in the fact that the control structure (which forms the basis of the analysis) can be created at any level of abstraction and with as much detail as required. Figure 7.7 depicts a typical control structure of a sociotechnical system at the highest level of abstraction. It captures the influence of legislations on system development and system operations. Each level in Figure 7.7 can be further detailed into nested control loops with control actions and feedbacks.

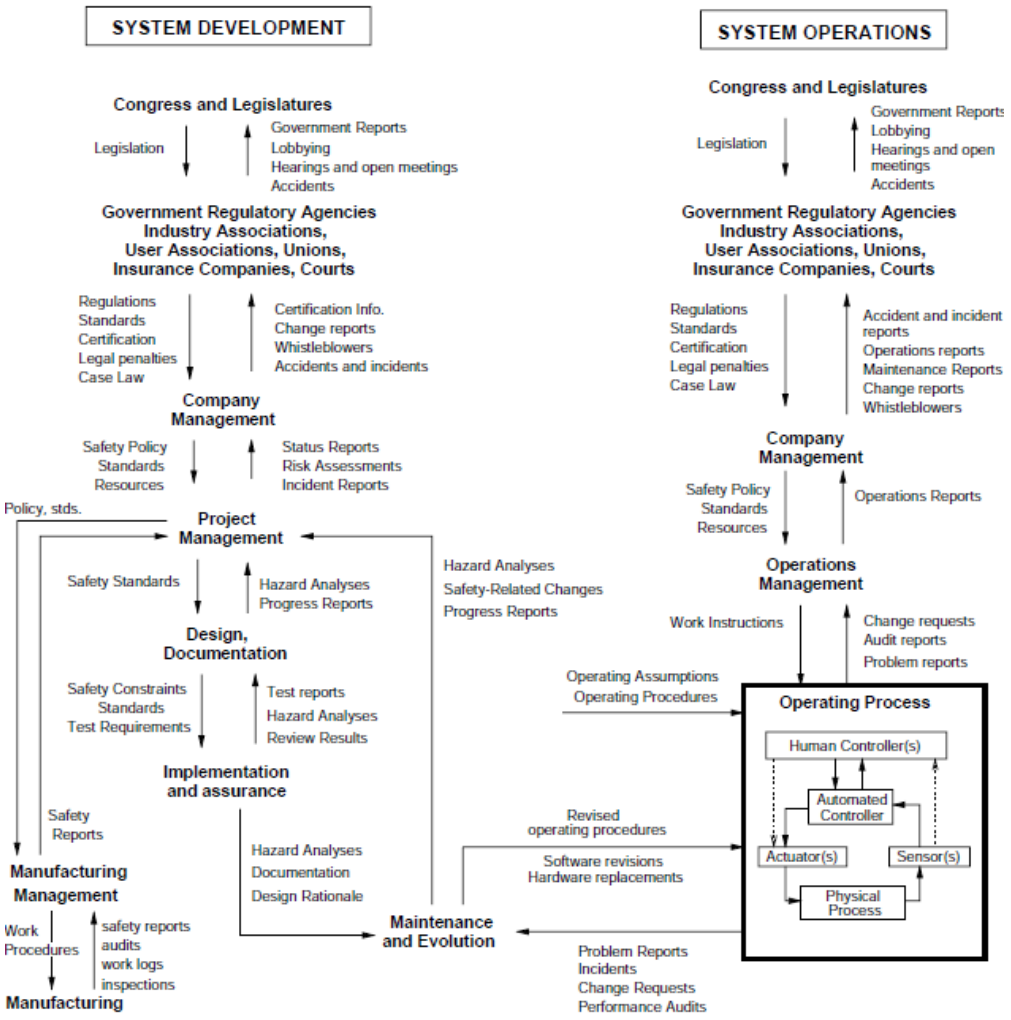

Figure 7.7: Sociotechnical control structure (Leveson, 2012)

7.2.5.3. STPA Step 3: Identification of UCAs

After the systems' (of interest) system level hazards / losses are identified (in step 1), and the system control structure is defined (in step 2), STPA Step 3 can be performed. STPA Step 3 identifies contextual deficiencies in control actions which can lead to a hazard or a loss. In order to identify these deficiencies, called as "Unsafe Control Action" (UCA), each control action (identified in the system control structure) is analysed in four general conditions:

- Not providing a control action causes a loss
- Providing a control action causes a loss
- Providing a control action too late, too early or out of sequence causes a loss
- Control action stopped too soon or applied too long causes a loss

In order to efficiently capture the above analysis to identify UCAs, a simple table (Table 7.4) can be used. Each element in the table is evaluated to against the system losses defined in step 1 (section 7.2.5.1), to determine whether it should be classified as a UCA. For instance, in the example discussed in Table 7.4, if the occupant doesn't provide the occupant key *"when the occupant doesn't want to get inside the pod"*, no loss is caused. On the other hand, when the context is changed to the occupant wanting to get inside the pod, not providing the occupant key will cause a loss, and is hence classified as a UCA. Each of the UCAs can then be separately analysed to create system safety requirements or identify missing requirements. Additionally, a UCA has a defined structure to it, i.e., actor, control action type, control action and context (Figure 7.8). It is important to maintain this structure as the proposed extension for test scenario generation (section 7.3) makes use of this UCA structure.

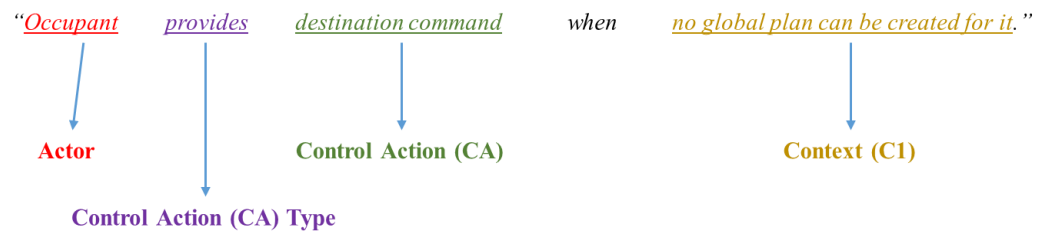

Figure 7.8: Structure of an Unsafe Control Action (UCA)

Table 7.4: Table for identifying Unsafe Control Actions (UCAs) with example UCAs

Control Action	**Not Providing causes a loss**	**Providing causes a loss**	**Too early, too late, out of sequence causes a loss**	**Stopped too soon or applied too long causes a loss**
Occupant Key	*Occupant doesn't provide occupant key when the occupant wants to get inside the pod." [L2]*	*"Occupant provides an incorrect key when the authorizer key is also incorrect and the remote operator key is also incorrect. - [L2]*	*Occupant provides the Correct Occupant key outside the valid slot of the authorizer key and/or remote key. - [L2]*	*Occupant provides occupant key for less than x seconds required to get into the vehicle when the person wants to get inside the pod. - [L2]*
Destination command	*Occupants doesn't provide destination command when they want to undertake a journey. - [L2]*	*Occupant provides destination command when no global plan can be created for it. - [L2]*		

Localized position	*Localizer doesn't provide localized position when vehicle is moving as a part of global plan. - [L1]*	*Localizer provides localized position with high covariance error. - [L2]* *Localizer provides the incorrect localized position when there is low covariance error. - [L2, L3]*	*Localizer provides localized position which corresponds to old real position due to latency. - [L1] [L2]*	*Localizer stops providing localized position while the vehicle is moving. - [L1] [L2]*

In Table 7.4, each UCA has been linked to its corresponding loss. It is important to note that UCAs identified in the table are only those which after analysis of a control action in either of the four general conditions lead to a loss as identified in step 1 of STPA. Linking UCAs to their corresponding loss also provides a degree of traceability in the analysis.

7.2.5.4. STPA Step 4: Identifying why UCAs might occur

After identifying Unsafe Control Actions (UCAs), STPA Step 4 involves identification of the causal reasons for the UCAs by analysing each control loop for each control action in the control structure created in step 2 (section 7.2.5.2). Causal scenarios are not fault states as identified in FMEAs. In an FMEA a known failure is analysed to understand the downstream effect and the question asked is "what failed". However, in STPA step 4, the question asked is "why it occurred". This can be explained by the fact that STPA offers a preventive outlook to failures whereas other methods like FMEAs look to mitigate the effect of the failure rather than to prevent it. For every UCA, two types of reasons must be considered (Leveson and Thomas, 2018) (Figure 7.9):

- Why would Unsafe Control Actions occur? (type a)
- Why would control actions be improperly executed or not executed, leading to hazards? (type b)

The two types of reasons for loss can be further elaborated with a detailed analysis by asking "why". A generic control loop is illustrated in Figure 7.10 which includes a controller, actuator, sensor and controlled process blocks. The detailed analysis of the scenarios can thus be split in four themes:

1. Inappropriate decisions (type a causal reason)
2. Inadequate feedback and other inputs (type a causal reason)
3. Inadequate control execution (type b causal reason)

4. Inadequate process behaviour (type b causal reason)

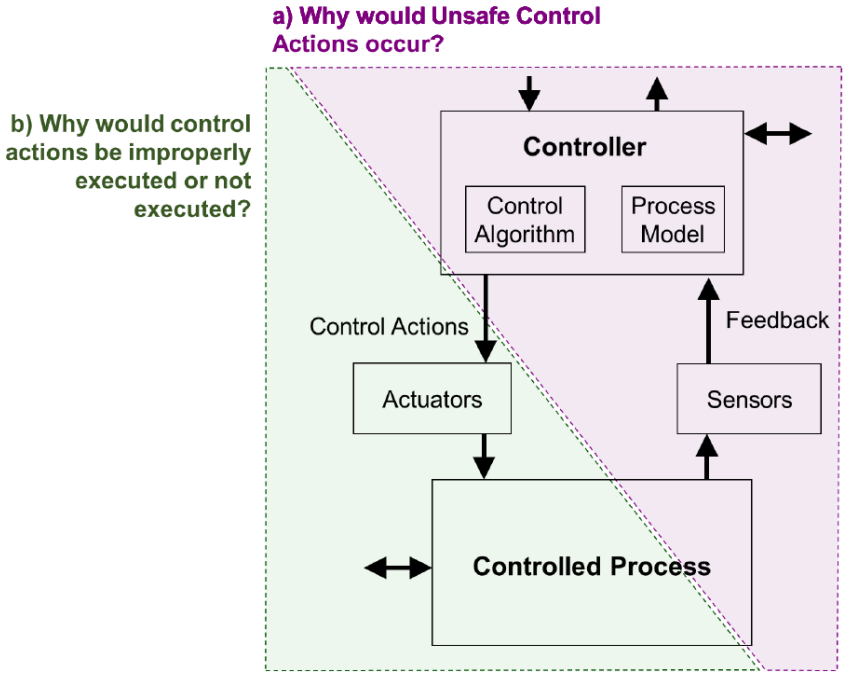

Figure 7.9: Two types of scenarios that must considered ((Leveson and Thomas, 2018) - page 43)

Inappropriate decisions can be made by the controller due to issues either with the 1) controller microprocessor 2) control algorithm or 3) the process model. Poorly designed or errors in control algorithms can lead to the inappropriate control action commanded by the controller. Also, the control algorithms make decisions based on the process model (including that of the world) it has. If the process model is invalid or obsolete, the decisions made by the control algorithm will inevitably be inadequate. Inadequate feedback can be of various types such as 1) loss of information due to input sampling frequency of the sensors or 2) due to conflicting feedback from different sensors (as experienced by the pilots in Air France 447 accident) or 3) incorrect feedback from the sensors. Inadequate control execution on the part of the actuators could be due to actuators receiving conflicting commands from different controllers. A classic case of such a situation would be a system in which both Adaptive Cruise Control (ACC) and Automatic Emergency Braking (AEB) are sending different brake commands to the brake actuators in a car. Other reasons for inadequate control in the control system include, if the actuators do not execute the command or if the actuator response time is too large. Similarly, inadequate process behaviour could be caused due to a various reasons. These include 1) more than one actuator

is trying to affect the controlled process or 2) if the controlled process doesn't react to the actuator inputs or 3) if the controlled process doesn't execute the control action correctly in response to the actuator input or 4) if the controlled process is missing some process elements required for its proper functioning.

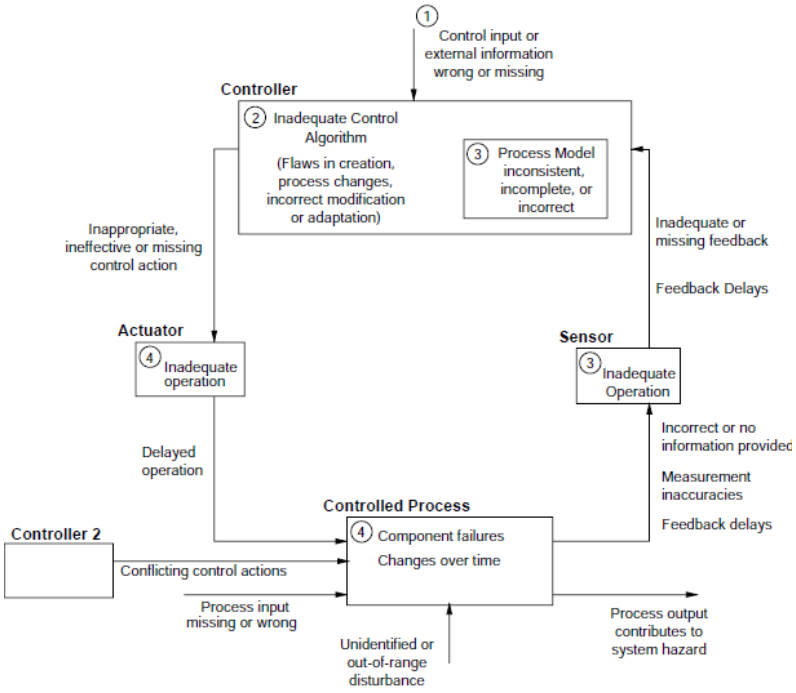

Figure 7.10: Classification of control flaws in a control loop (Leveson, 2012)

In order to efficiently capture the above analysis to identify UCAs, a simple table (

Table 7.5) can be used. For a UCA to occur, the process model has a belief which is triggering a UCA from the control algorithm (assuming the control algorithm is correct). Thus, the first aspect to understand is the belief of the process model. Next, the reason(s) for this belief of the process model are identified. Finally, the situations for the corresponding reasons are identified or for control action being not followed as desired.

Table 7.5 illustrates the Step 4 analysis for one UCA.

Table 7.5: Table for conducting STPA Step 4

Unsafe Control Action (UCA)	**Process Model: believes (B1)**	**Process Model: believes that because (B2))**	**Potential control action not followed / How could this happen**
Occupant doesn't provide occupant key when the occupant wants to get inside the pod. - (L2)	*Occupant believes occupant has provided occupant key.*	*Occupant believes that because the HMI display says key has been provided.* *Occupant believes that because there is no confirmation feedback provided to the occupant*	*HMI display's algorithm is faulty.* *HMI feedback absent in design.*
	Occupant believes occupant does not need to provide a key.	*Occupant believes that because they are unfamiliar with the technology.*	*Incorrect marketing done.* *Lack of labelling outside the vehicle.*

7.2.5.5. STPA: Advantages and differentiators

Unlike other methods like FMEA, FTA, ETA which are focussed on identifying a downstream effects of failures or chain of events, STPA enables the analyst to understand the reason behind the failures and identifies requirements to prevent the failure from re-occurring. One of the significant benefits of using STPA is its capacity to identify diverse causal factors (component failures, component interactions, requirements flaws, human-errors, design flaws, societal issues, organizational issues etc.) (France, 2017; Leveson, 2012).

While STPA has been used widely in aviation (Fleming et al., 2013), space industry (Ishimatsu et al., 2010), military applications, organisational studies, its application in the automotive domain has been comparitively limited. This is partly because there are other established methods (e.g. FMEA, FTA, HAZOP etc.) with their corresponding software tools which make applying them much less time consuming. While some attempts of formalising the STPA analysis have been made (Thomas et al., 2015), the process is still elaborate and time consuming. With the absence of software tools to aid the analysis, the process can be commercially expensive. However, as ADAS and ADS are complex systems in which hazards can occur due to a number of reasons (mentioned earlier), STPA offers the

most complete set of hazards (Sotomayor Martínez, 2015). Thus, STPA was chosen as the method for hazard identification which stage one of hazard based testing.

7.3. Extending STPA to create test scenarios

In this section, a methodology to create test scenarios from the STPA output (step 1-4) is introduced. This involves identifying parameters for parametrisation of various elements of STPA outputs. First, it is important to understand how a test scenario needs to be constructed, or in order words what are the components of a test scenario. It is suggested that a test scenario should consist of a world, actors and their behaviour (Ulbrich et al., 2015). However, it is the author's understanding (based on prior-industrial experience in the automotive industry) that a complete test scenario definition should also contain the *"pass criteria"* for the corresponding test scenario. For a test scenario to be usable for testing, it is essential to know the criteria when the scenario is considered to be a pass. With this understanding, a complete test scenario description will have four components:

1. Scenery
2. Dynamic elements
3. Pass criteria
4. Additional context

Scenery defines all geo-spatially stationary objects in the Operational Design Domain (ODD) of the vehicle (Ulbrich et al., 2015). It also includes environmental conditions such as weather, visibility etc., along with road layouts, road furniture (e.g. barricades), traffic lights etc. Thus, scenery can be sub-categorised into various parameters (with their sub-parameters). However, selection of the scenery to be used for testing is influenced by the ODD of the Subject Under Test (SUT). While urban roads or city centres or motorway roads are part of scenery elements, if the ODD of a LSAD system is urban roads and city centres, then the scenery parameters will be selected accordingly.

All moving objects and actors in the world comprise the dynamic elements category (Ulbrich et al., 2015). Dynamic elements are sub-categorised into various parameters. Like scenery elements, selection of dynamic element parameters is also dependent on the ODD of the SUT. For example, if the ODD of a system is only motorways, then pedestrians generally will not expected to be part of the dynamic elements on the motorway (as pedestrians are not allowed on motorways. Figure 7.11 illustrates the top-level scenery element parametrisation. At the top-level, scenery can be classified into "map" and "environmental" parameters. Figure 7.12 illustrates the top-level dynamic element parametrisation. At the top-level, dynamic elements can be classified into scripted traffic and

non-scripted traffic. Scripted traffic refers to non-SUT agents which have a pre-defined path in the scenario. Non-scripted traffic refers to agents which may be part of a traffic model or an actor model without predefine path.

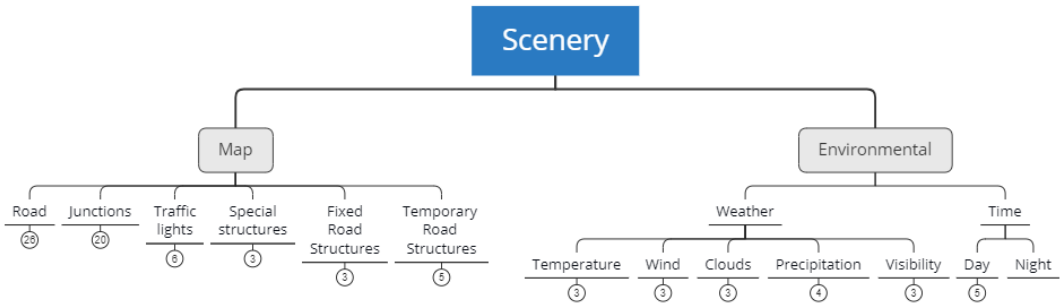

Figure 7.11: Scenery element parametrisation (top level)

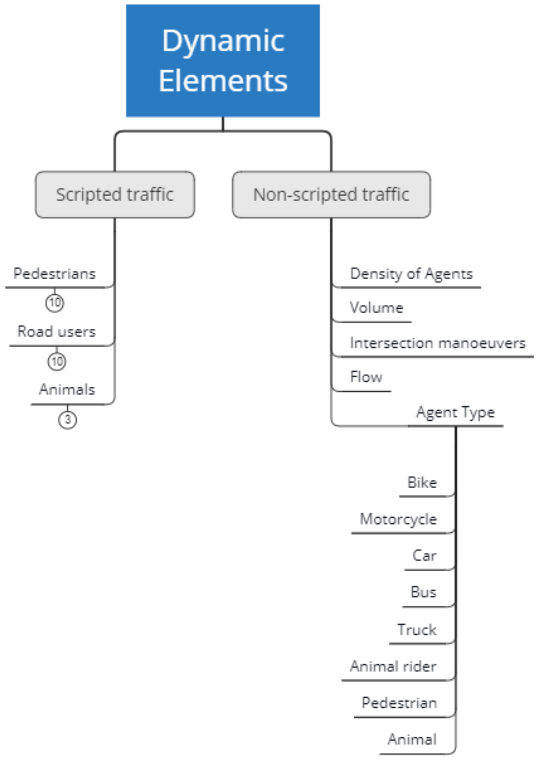

Figure 7.12: Dynamic element parametrisation (top level)

The scenery and the dynamic elements for the test scenario are selected from the library based on the defined Operational Design Domain (ODD) of the SUT (in STPA Step 1). The scenery and the dynamic elements together form the base world for the test scenario and

their corresponding parameters depending upon the ODD form the base parameters for a scenario. The pass criterion defines the set of conditions (internal to SUT or external to SUT) for which the test scenario will be considered as a pass. Additional context refers to the context element of the UCA (STPA step three) and the reasons for *"Potential control action not followed / How could this happen"* in STPA step four.

7.3.1.1. Additional Context scenario parameters

As mentioned in section 7.2.5.2, an Unsafe Control Action (UCA) is structured into four components. As an example, let us consider the following UCA (Figure 7.13).

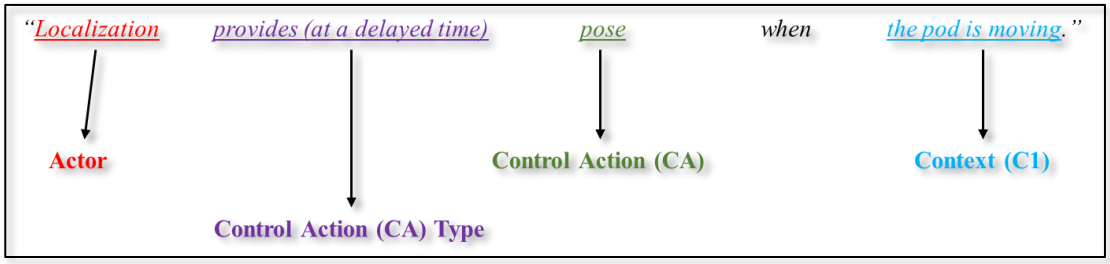

Figure 7.13: Example 1: Unsafe Control Action from STPA of Low-Speed Automated Driving system [UCA# 8c]

Scenario parameters are identified for the scenery, dynamic elements (both as per the ODD defined in STPA step one) and the additional context. The additional context comes from both STPA step three (from the UCA) and STPA step four.

One of the parameters identified for the additional context is the "context (C1)" of a UCA (Figure 7.13). In the example in Figure 7.13, the content element is *"...when the pod is moving"*. The corresponding parametrisation of *"pod is moving"* is done with respect to trajectory and speed, which form the additional context.

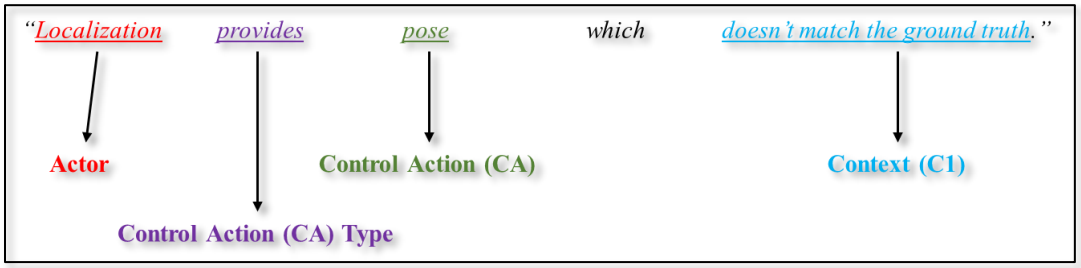

Figure 7.14: Example 2: Unsafe Control Action from STPA of Low-Speed Automated Driving system [UCA# 8b]

In another UCA example (Figure 7.14), the *"context element"* of the UCA is: *"doesn't match the ground truth"*. The corresponding parameter for *"ground truth"* is the base map, i.e., scenery element. In this case, the base map is a STPA specific parameter (in addition to creating the base world). The difference between an STPA specific parameter and a base

parameter is that in the execution of test cases for the test scenario, the STPA specific parameters will have a higher resolution when parameter values are chosen and base parameters will have a lower resolution (or random selection depending on the defined ODD).

The second part of the "additional context" comes from STPA step four analysis. Table 7.6 illustrates the STPA step four analysis identifying causal factors for UCA# 8c and UCA# 8b (Figure 7.13 and Figure 7.14).

Table 7.6: STPA Step 4 analysis for LSAD system for UCA# 8b and 8c

Unsafe Control Action (UCA)	**Process Model: believes (B1)**	**Process Model: believes that because (B2))**	**Potential control action not followed / How could this happen (B3)**
[UCA# 8b] *Localization provides pose which doesn't match the ground truth. - [L2][L1]*	*Obtaining Pose block believes it has the correct pose.*	*Obtaining Pose block believes this because the Covariance Error is low (i.e., sensor data is coherent. (But actually sensor data is incorrect)*	*This could be because sensor feed is delayed in time.*
[UCA# 8c] *Localization provides pose at a delayed time when the pod is moving. – [L1]*	*Obtaining Pose block believes there is no delay.*	*Obtaining Pose block believe this because the data packets it receives matches the current time and subsequently stamps the sent data to the current time.*	*This could be due to scheduling of the Obtaining pose block being delayed.* *This could be due to algorithm execution taking longer than expected, therefore the output data is based on processing of old data.* *This could be due to internal communications error (Bus overload or Bus packet data corruption)*

For the second part of the additional context, the *"Potential control action not followed / how the situation could happen"* B3 element of STPA step four is parametrised. E.g. if the *"how"* element is: *"sensor feed was delayed"* [UCA# 8b], then the parametrisation of *"type of sensor feed"* and *"delay time"* is done. UCA# 8c has various *"how"* elements. For the reason – *"... due to scheduling of the Obtaining pose block being delayed"*, the parametrisation is done for the delay time. For the reason – *"Bus overload"*, the parametrisation is done for the *"amount of overload"*.

7.3.1.2. Pass / fail criteria for the scenario

The pass criteria are identified from STPA step four. The aim of the STPA process is to identify safety requirements to prevent a UCA from happening (if possible). Consider a control loop (e.g. Figure 7.15), for a UCA to occur, the process model of the controller has a belief based on which it believes that the control action it is directing is safe (when actually it is unsafe, i.e. a UCA). Let us call this belief as B1. If B1 were not true, the controller would not direct the original control action (i.e., original UCA). Therefore, one of the pass criterion would be defined as the negation of the belief B1 (process model belief), i.e., B1´, as identified in STPA step four. Secondly, the process model has the belief (B1) because of some reasons. Let us call these reasons causing the process model belief B1 as B2. Once again, if these reasons B2 were not true, B1 would not be true (when treated recursively) and subsequently, the controller would not direct the UCA. Thus, the second pass criterion is the negation of the reasons for the process model belief, i.e., B2´. Thus, the two pass criteria coming out of STPA step four for each UCA are B1´ and B2´. Thus, the pass criteria for the scenarios is the negation of the *"process model belief"* and negation of the *"reason for the process model belief"*. It is possible that there could be multiple process model beliefs causing the UCA and multiple corresponding reasons for those beliefs. In such a situation, there will be more than one test scenario for a single UCA and each test scenario will have its corresponding pass criteria. Table 7.7 illustrates the pass criteria based on UCA# 8b.

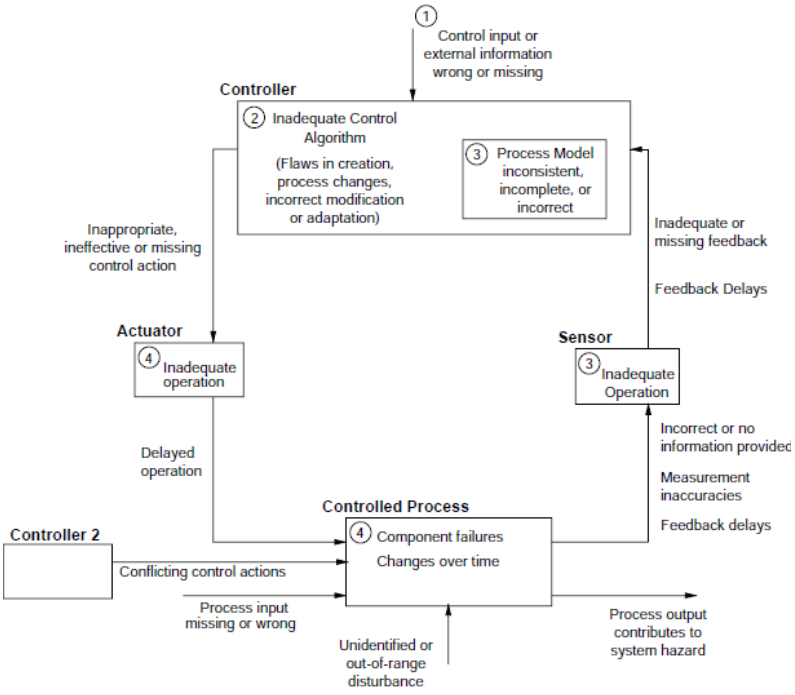

Figure 7.15: Generic control loop

Table 7.7: Pass criteria based on process model belief and reasons for the process model belief

Unsafe Control Action (UCA)	**Process Model: belief and its reasons**	**Pass criteria**
[UCA# 8b] Localization provides pose which doesn't match the ground truth. - [L2][L1]	B1 *Obtaining Pose block believes it has the correct pose.*	B1' *Obtaining Pose block shall believe it has the INCORRECT pose.*
	B2 *Obtaining Pose block believes this because the Covariance Error is low (i.e., sensor data is coherent. (But actually sensor data is incorrect)*	B2' *Covariance Error shall be HIGH.*

The scenario description (based on UCA# 8b) can also be written in the following notation (Table 7.8):

Table 7.8: Scenario based on UCA# 8b

Scenery	Urban areas
Dynamic elements	Pedestrians, vehicles (random selection)
Context	Base map (different types) Type of sensor feed Sensor feed delay time
Pass criteria (PC)	**PC1:** Obtaining Pose block shall believe it has the INCORRECT pose. **PC2:** Covariance Error shall be HIGH
Number of Parameters	3
Number of Scenarios	$({}^3C_1 + {}^3C_2 + {}^3C_3 = 7)$ x 2 (number of pass criteria) = 14 or $(2^k - 1)$ x 2, where k is the number of parameters (for corresponding B1` and B2`)

The scenario described in Table 7.8, involves adding delay in various sensor(s) feed(s). These scenarios evaluate if the covariance error detects the mismatch between the ground truth and the current pose calculated based on the sensor feed (which has been delayed).

7.4. Applying proposed test scenario generation method to a Low-Speed Automated Driving System: A Real-World Case Study

In order to evaluate the applicability of the proposed method to create test scenarios, the method was applied to a real-world automated driving system as a case study. In this case-study, Aurrigo Technology's fully automated Low-Speed Automated Driving (LSAD) system similar to Figure 7.16, was the system under consideration or the system under test (SUT). As discussed in section 7.2.5, in order to apply the proposed method to create test scenarios using Hazard Based Testing, first STPA of the LSAD system was conducted.

Figure 7.16: Low-Speed Automated Driving system (pod)
(image courtesy: Aurrigo Technology)

7.4.1. STPA Step 1: LSAD system

The first step involves defining the system and the losses. The Aurrigo LSAD system is an SAE Level 4 system, i.e., it is fully autonomous in a dedicated Operational Design Domain (ODD). The dedicated ODD for the LSAD system was its predefined routes in an urban environment. Additionally, the larger mobility system (of which each LSAD system was a

part of) consisted of a dispatcher, web-server and a fleet supervisor. The LSAD system equipped vehicle had electric propulsion with break-by-wire and steer-by-wire functionality. It had a diverse range of sensors including multiple LiDAR and Cameras as a part of its sensor suite. For the Aurrigo LSAD system the following were identified as losses:

1. Crash with a static object or a dynamic actor (L1)
2. Not completing the journey with passenger and cargo (L2)
3. Time of journey being too long, i.e., service target not met (L3)

Loss 2 and Loss 3 were defined as losses as the Aurrigo LSAD system is a part of a larger mobility service system. An incomplete journey or if a journey takes too long would lead to loss of customer satisfaction which will ultimately lead to loss of revenue for the business (mobility service). This is an important aspect of STPA as it can capture analysis for social-technical losses also.

7.4.2. STPA Step 2: LSAD system

STPA step 2 involves creating a control structure of the system under test (SUT), i.e., Aurrigo's LSAD system. One of the most important aspect of the control structure development is the identification of the interactions (control commands and feedback) between the subsystems. The LSAD control structure can be classified into five different abstractions. The first abstraction consists of the *"world"* in which the LSAD is being deployed. The second abstraction is the *"raw sensing"* of the world in which the LSAD is being deployed. The third abstraction is the *"autonomous control system"* of the LSAD. The forth abstraction being the *"LSAD actuation system"* which responds to the autonomous control system to make the LSAD to move in the world. The fifth abstraction is the *"human input"* to the LSAD.

In order to create the control structure for the pod system, first the various sub-systems needed to be identified. These subsystems are:

- Customer
- Sensors
- Daily Authorizer
- World
- Autonomous Control System (ACS)
    - Remote Operator
    - SW Authorizer
    - Localization
    - Obstacle detection classifier

    - Global path planner
    - Local path planner
    - Kinematic models
- LSAD Actuation
    - Vehicle Management system
    - Steering sensor estimator
    - Braking pressure estimator
    - Motor torque estimator

The control commands and feedback interaction between the LSAD sub-systems is captured in Figure 7.17. In Figure 7.17, red arrows indicate control actions and green arrows indicate feedback actions. The STPA analysis carried out as a part of this thesis involved the sensors, fleet supervisor, world and the autonomous control systems subsystems. The analysis discussed in this section is for a part of the autonomous control system as depicted in the highlighted region in Figure 7.18.

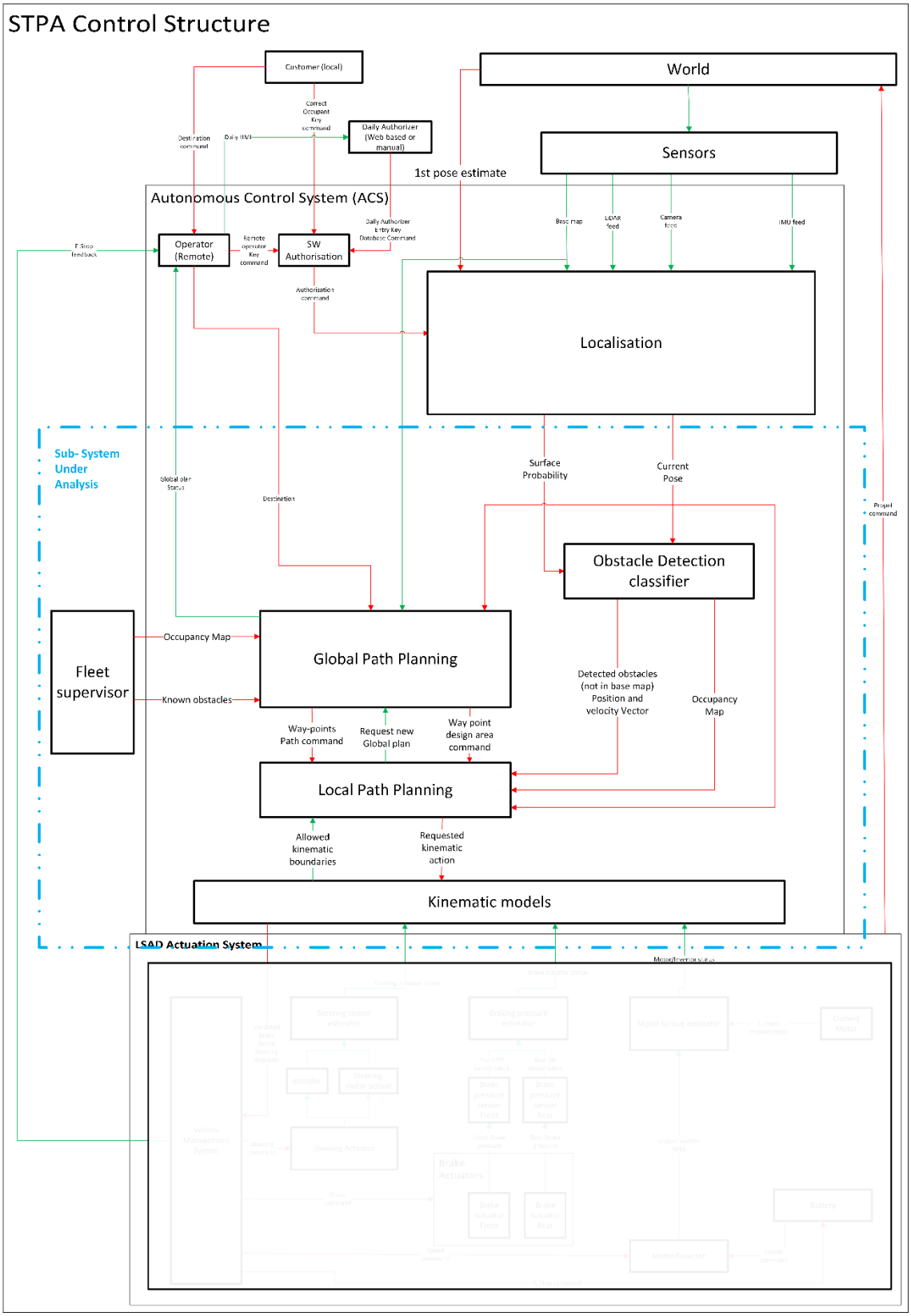

Figure 7.17: STPA control structure for LSAD system

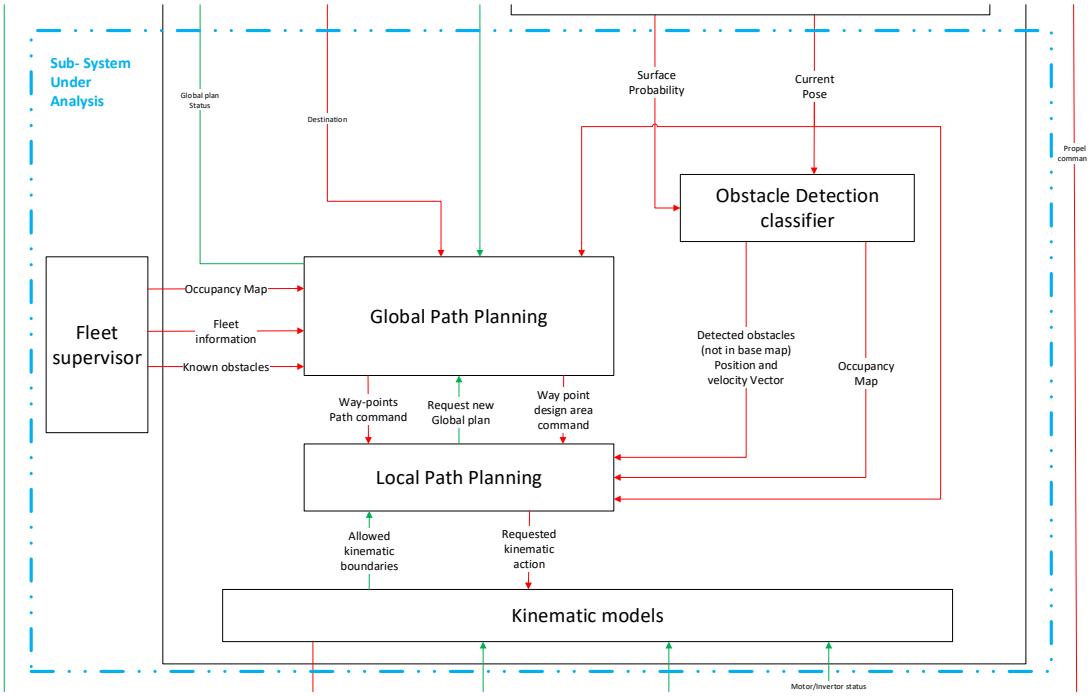

Figure 7.18: Highlighted Region of the LSAD system control structure

7.4.3. STPA Step 3: LSAD system

STPA step three involves analysing each control action to identify Unsafe Control Actions (UCAs). For the autonomous control system part of the LSAD system, analysis of 20 control actions (in Figure 7.17) resulted in 107 UCAs. For the purpose of this thesis (and due to confidentially reasons), only the UCAs for the control actions (highlighted in Figure 7.18) are discussed. Some of the identified UCAs are illustrated in Table 7.9.

For the nine control actions identified in Figure 7.18, 45 Unsafe Control Actions were identified, with eleven UCAs being associated with a single control action – *"requested kinematic command"*. Other control actions like *"occupancy map"* had nine UCAs associated with it, while control actions *"surface probability"* and *"fleet information"* had five and seven UCAs associated with them respectively.

Table 7.9: UCA table with some of the UCAs identified for the LSAD system

Control Action	**Not Providing causes a loss**	**Providing causes a loss**	**Too early, too late, out of sequence causes a loss**	**Stopped too soon or applied too long causes a loss**
Destination command	[UCA# 4a] *Occupant doesn't provide destination command when occupant wants to undertake a journey. - [L2]*	[UCA# 4b] *Occupant provides destination command when no global plan can be created for it. - [L2]*	-	-
Current pose (position)	[UCA# 8a] *Localization doesn't provide updated pose when the vehicle is requested to move. - [L1][L2]*	[UCA# 8b] *Localization provides pose which doesn't match the ground truth. – [L1][L2]*	[UCA# 8c] *Localization provides pose at a delayed time when the pod is moving. - [L1] [L2]*	[UCA# 8d] *Localizer provides the current position for too long when vehicle in moving. - [L1] [L2]*
Surface probability	[UCA# 10a] *DPM doesn't provide Surface probability when the pod is requested to move. – [L1]*	[UCA# 10b.1] *DPM provides incorrect Surface probability when the Rolling velocity vector is correct. [L2]* [UCA# 10b.2] *DPM provide incorrect Surface probability when Odometry is also incorrect. [L1]*	[UCA# 10c] *DPM provides correct surface probability but delayed while pod is moving. – [L1]*	[UCA# 10d] *DPM provides same surface probability for more than 10ms while the pod is moving. – [L1]*
Detected obstacles vector	[UCA# 12a] *Classifier doesn't provide Detected obstacles vector when obstacle is in vehicle trajectory. – [L1]*	[UCA# 12b] *Classifier provides Detected obstacles vector when obstacle is not vehicle trajectory. – [L2] [L3]*	[UCA# 12c] *Classifier provides Detected obstacles vector x seconds delayed when obstacle is vehicle trajectory. – [L1]*	-
Way Points design area command	[UCA# 14a] *GPP doesn't provide design area when there are obstacles on the path. – [L2]*	[UCA# 14b] *GPP provide large design area when there are obstacles on the path and when design area is larger than operational design domain. – [L2] [L3]*	[UCA# 14c] *GPP provides design area too early or late when design area is larger than true ODD in real time. – [L1, L2]*	[UCA# 14d] *GPP stops providing design area when vehicle is moving and there is obstacle along the vehicle trajectory. – [L2]*

7.4.4. STPA Step 4: LSAD system

After identifying the UCAs (in the previous section), the reasons for the occurrence of the UCAs were identified as part of STPA Step four. Table 7.10 illustrates the causal reasons for some of the UCAs.

Table 7.10: UCA Step 4 table for some of the UCAs for the LSAD system

Unsafe Control Action (UCA)	**Process Model: believes (B1)**	**Process Model: believes that because (B2))**	**Potential control action not followed / How could this happen (B3)**
[UCA# 4a] *Occupants don't provide destination command when they want to undertake a journey.* – [L2]	*Occupants believe they have provided the key.*	*Occupants believe that because the HMI display says key has been provided.* *Occupants believe that because there is no confirmation feedback provided to the occupant*	*HMI display's algorithm is faulty.* *HMI feedback absent in design.*
	Occupant believe they don't need to provide a key.	*Occupants believe that because they are unfamiliar with the technology.*	*Incorrect marketing done.* *Lack of labelling outside the vehicle.*
[UCA# 4b] *Occupant provides destination command when no global plan can be created for it.* – [L2]	*Occupant believes it is possible to create a global plan.*	*Occupant believes that because the system has told implicitly (by acceptance of booking) or explicitly that it is possible.*	*This could be because on the day of the trip there are not enough vehicles or supervisors or battery in the pods is not available.* *This could be because of an incorrect base map.*
[UCA# 8a] *Localization doesn't provide updated pose when the vehicle is requested to move.* - [L1] [L2]	*Obtaining Pose block believes there is large Covariance error.*	*Obtaining Pose blocked believes this because the sensor data provided is incoherent.*	*This could be because processing of the internal filter is incorrect.*
	Obtaining Pose believes it is providing.	*Obtaining pose block believes that because Covariance Error is low.*	*This could be due to internal communications error (Bus overload or Bus packet data corruption)*

[UCA# 8b] *Localization provides pose which doesn't match the ground truth.* - [L2][L1]	*Obtaining Pose block believes it has the correct pose.*	*Obtaining Pose block believes this because the Covariance Error is low (i.e., sensor data is coherent. (But actually sensor data is incorrect)*	*This could be because sensor feed is delayed in time.*
[UCA# 8c] *Localization provides pose at a delayed time when the pod is moving.* – [L1]	*Obtaining Pose block believes there is no delay.*	*Obtaining Pose block believe this because the data packets it receives matches the current time and subsequently stamps the sent data to the current time.*	*This could be due to scheduling of the Obtaining pose block being delayed.* *This could be due to algorithm execution taking longer than expected, therefore the output data is based on processing of old data.* *This could be due to internal communications error (Bus overload or Bus packet data corruption)*
[UCA# 8d] *Localizer provides the current position for too long when vehicle in moving.* – [L1][L2]	*Obtaining Pose block believes the incoming sensor data shows no change in position.*	*Obtaining pose block believes that because surface probability command doesn't match rolling velocity vector and the Covariance Error is below the threshold.*	*This could be because of a featureless or less features in environment (i.e., repetitive or empty - e.g. tunnel, airfield, ware-house bays)*
	Obtaining Pose block believes it is sending the updated position.	*Obtaining pose block believes because it generates the data packets with the correct time stamp.*	*This could be due to scheduling of the Obtaining pose block being delayed.* *This could be due to algorithm execution taking longer than expected, therefore the output data is based on processing of old data.* *This could be due to internal communications error (Bus overload or Bus packet data corruption)*
[UCA# 10a] *DPM doesn't provide Surface probability when the pod is requested to move.* – [L1]	*DPM believes it is providing the surface probability.*	*DPM believes that because it is generating the packets and trying to send them.*	*This could be due to internal communications error (Bus overload or Bus packet data corruption).*

[UCA# 10b.1] *DPM provides incorrect Surface probability when the Rolling velocity vector is correct.* - [L2]	*DPM believes it correctly quantified the surface probability when the all the sensor data was available and coherent.*	*DPM believes that because the covariance error is low and sensor data is correct.* *DPM believes that because the covariance error is low but sensor data is incorrect.* *DPM believes that because the matching algorithm told it so.*	*This could be because the base map is incorrect and doesn't match with ground truth.* *This could be because the environment is not feature rich (i.e. repetitive or empty e.g. tunnel, airfield, ware house).* *This could be because the matching algorithm is incorrect.*
[UCA# 10b.2] *DPM provides incorrect Surface probability when Odometry is also incorrect and both are coherent.* - [L1]	*DPM believes it correctly quantified the surface probability when the all the sensor data was available and coherent*	*DPM believes that because Covariance Error is low.*	*This could be due to internal communications error (Bus overload or Bus packet data corruption).* *This could be due to DPM or OPM algorithm being incorrect.* *This could be due to a malicious attack.*
[UCA# 10c] *DPM provides correct surface probability but delayed while pod is moving.* – [L1]	*DPM believes it correctly quantified the surface probability when the all the sensor data was available and coherent*	*DPM believes that because the matching algorithm was able to match the sensor data to the base match.*	*This could be due to internal communications error (Bus overload or Bus packet data corruption).* *This could be because the matching algorithm is incorrect.*
		DPM believes that because the Covariance Error is low.	*This could be due to delay in sensor(s) feed(s).* *This could be due to a malicious attack.*
[UCA# 10d] *DPM provides same surface probability for more than 10ms while the pod is moving.* – [L1]	*DPM believes that it is not providing the same surface probability*	*DPM believes that because it is able to match the sensor data to the base match*	*This could be because of the sensor output is stuck at a particular value.* *This could be because the matching algorithm is incorrect.* *This could be due to internal communications error (Bus overload or Bus packet data corruption).* *This could be due to a malicious attack.*

	DPM believes it is providing the correct surface probability based on the input sensor data	*DPM believes that because the matching algorithm was able to match the sensor data to the base match.*	*This could be because the environment is not feature rich (i.e. repetitive or empty e.g. tunnel, airfield, ware house).* *This could be due to base map being incorrect.* *This could be due to internal communications error (Bus overload or Bus packet data corruption).* *This could be due to a malicious attack.*

7.4.5. Creating test scenarios and scenario parameters

Once the UCAs and their corresponding causal reasons are identified, the next step involves parametrisation for scenario creation. As discussed in section 7.3, a test scenario comprises of scenery elements, dynamic elements, context and pass criteria. The base parameters are selected depending upon the ODD specification of the LSAD. In this case study, the LSAD's ODD included urban areas with pre-determined routes only.

Parametrisation for scenario generation involves parametrisation of the context element of the UCA and the parametrisation of the *"Potential control action not followed / How could this happen"* element. In this section, some of the UCAs identified in Table 7.9 with their corresponding causal reasons identified in Table 7.10 are parametrised for scenario creation. Additionally, the corresponding pass criteria are also identified by negating the process model belief and reasons for the belief as identified in Table 7.10. Table 7.11 captures the scenario parameters and their pass criteria. The scenario representation of the Table 7.11 could also be made similar to Table 7.8.

Based on the parameters identified for each of the scenarios, test cases can be created by assigning values to each of the parameters. While the parameter value selection process is out of scope of this thesis, some of the potential approaches are discussed in section 7.5.

Table 7.11: Scenarios based on STPA analysis of the LSAD system

UCA	Potential control action not followed / How could this happen (B3)	UCA Specific context	UCA context parameters	UCA how parameters	Pass Criterion 1	Pass Criterion 2
[UCA# 8a] *Localization doesn't provide updated pose when the vehicle is requested to move. [L1] [L2]*	*This could be because processing of the internal filter is incorrect*	*when the vehicle is requested to move*	*Velocity*	*Error in internal filter*	*Obtaining Pose block believes there is small Covariance error.*	*Sensor data provided is coherent.*
[UCA# 8b] *Localization provides pose which doesn't match the ground truth. - [L2][L1]*	*This could be because sensor feed is delayed in time.*	*which doesn't match the ground truth*	*Scenery (map data)*	*Delay time* *Type of sensor delayed*	*Obtaining Pose block shall not believe it has the correct pose.*	*Covariance Error shall be is high.*
[UCA# 8c] *Localization provides pose at a delayed time when the pod is moving. – [L1]*	*This could be due to scheduling of the Obtaining pose block being delayed.* *This could be due to algorithm execution taking longer than expected, therefore the output data is based on processing of old data.*	*when the pod is moving*	Velocity	Delay time Algorithm execution time duration	*Obtaining Pose block shall believe there is delay in pose signal.*	*Obtaining Pose block shall believe that the data packets it receives don't match the current time.*

	This could be due to internal communications error (Bus overload or Bus packet data corruption)			Type of communication error		
[UCA# 8d] *Localizer provides the current position for too long when the pod in moving.* – [L1][L2]	*This could be because of a featureless or less features in environment (i.e., repetitive or empty - e.g. tunnel, airfield, ware-house bays)*	*when the pod in moving*	Velocity	Number of features in base world	*Obtaining Pose block shall not believe the incoming sensor data shows no change in position*	*Surface probability command shall match rolling velocity vector and the Covariance Error shall be below the threshold.* *Or* *Surface probability command shall not match rolling velocity vector and the Covariance Error shall not be below the threshold.*
[UCA# 8d] *Localizer provides the current position for too long when the pod in moving.* – [L1][L2]	*This could be due to scheduling of the Obtaining pose block being delayed.* *This could be due to algorithm execution taking longer than expected, therefore the output data is based on processing of old data.* *This could be due to internal*	*when the pod in moving*	Velocity	Delay time Algorithm execution time duration	*Obtaining Pose block shall not believe it is sending the updated position.*	*Obtaining pose block shall not believe its data packets have correct time stamp.*

	communications error (Bus overload or Bus packet data corruption)			Type of communication error		
[UCA# 10a] *DPM doesn't provide Surface probability when the pod is requested to move. – [L1]*	*This could be due to internal communications error (Bus overload or Bus packet data corruption).*	*when the pod is requested to move*	Velocity	Type of communication error	*DPM shall not believe it is providing the surface probability.*	*DPM shall not believe it is sending data packets.*
[UCA# 10b1] *DPM provides incorrect Surface probability when the Rolling velocity vector is correct. [L2]*	*This could be because the base map is incorrect and doesn't match with ground truth.*	*when the Rolling velocity vector is correct.*	Rolling velocity vector	Type of base map Storage mechanism of base map (coordinate system, format etc.)	*DPM believes it has not correctly quantified the surface probability when the all the sensor data was available and coherent.*	*DPM shall not believe that because the covariance error is low and sensor data is correct.*
	This could be because the environment is not feature rich (i.e. repetitive or empty e.g. tunnel, airfield, ware house)	*when the Rolling velocity vector is correct.*	Rolling velocity vector	Number of features in base world		*Covariance error shall be high.*
	This could be because the matching algorithm is incorrect.	*when the Rolling velocity vector is correct.*	Rolling velocity vector	Type of errors in matching algorithm		*The matching algorithm shall not say that the surface probability is correctly quantified.*
[UCA# 10b2] *DPM provides incorrect Surface probability when Odometry is also incorrect and both (surface probability and odometry) are coherent. - [L1]*	*This could be due to internal communications error (Bus overload or Bus packet data corruption).* *This could be due to DPM algorithm being incorrect.* *This could be due to a malicious attack.*	*when Odometry is also incorrect and both (surface probability and odometry) are coherent*	Deviation in odometry data from ground truth Deviation in surface probability data from ground truth Coherency deviation between odometry and surface probability data	Type of communication error Type of errors in matching algorithm Type of malicious attack	*DPM shall not believe that it correctly quantified the surface probability.*	*DPM shall believe that because Covariance Error is high.*

[UCA# 10c] *DPM provides correct surface probability but delayed while pod is moving. – [L1, L3]*	*This could be due to internal communications error (Bus overload or Bus packet data corruption).* *This could be because the matching algorithm is incorrect.*	*while pod is moving*	Velocity	Type of communication error Type of errors in matching algorithm	*DPM shall not believe that it correctly quantified the surface probability.*	*Matching algorithm shall not be able to match the sensor data to the base map.*
	This could be due to delay in sensor(s) feed(s). *This could be due to a malicious attack.*			*Delay time* *Type of sensor delayed* Type of malicious attack		*DPM shall believe that the Covariance Error is high.*
UCA# 10d] *DPM provides same surface probability for more than 10ms while the pod is moving. – [L1, L3]*	*This could be because of the sensor output is stuck at a particular value.* *This could be because the matching algorithm is incorrect.* *This could be due to internal communications error (Bus overload or Bus packet data corruption).* *This could be due to a malicious attack.*	*while the pod is moving*	Velocity	*Type of sensor(s) whose output is frozen* *Freeze duration of sensor(s) feed* Type of errors in matching algorithm Type of communication error Type of malicious attack	*DPM believes that it is providing the same surface probability*	*DPM shall not be able to match the sensor data to the base map.*
	This could be because the environment is not feature rich (i.e. repetitive or empty e.g.			Number of features in base world	*DPM shall not believe it is providing the correct surface probability.*	*Matching algorithm shall not be able to match the sensor data to the base map.*

	tunnel, airfield, ware house). *This could be due to base map being incorrect.* *This could be due to internal communications error (Bus overload or Bus packet data corruption).* *This could be due to a malicious attack.*			Type of base map Storage mechanism of base map (coordinate system, format etc.) Type of communication error Type of malicious attack		

Table 7.8 illustrates the relation between number of parameters, pass criteria and number of generated scenarios. Applying the proposed extension method (discussed in section 7.3) to STPA of the LSAD system (Figure 7.18), for the nine control actions, 350 parameters were identified which lead to the creation of 3398 test scenarios. Table 7.12 illustrates the number scenarios corresponding to each of the UCAs associated with the nine control actions identified in Figure 7.18.

Table 7.12: Number of scenarios for each UCA

Control Action	UCA#	Number of Scenarios
Destination command	4a	28
	4b	30
Current Pose	8a	12
	8b	14
	8c	30
	8d	36
Surface Probability	10a	6
	10b1	18
	10b2	126
	10c	44
	10d	188
Detected Obstacles Vector	12a	104
	12b	260
	12c	126
Way Points Path Command	13a	206
	13b1	62
	13b2	14
Way Points Design Area Command	14a	28
	14b	14
	14c	6
	14d	50
Requested Kinematic Command	15a	76
	15b1	76
	15b2	30
	15b3	74
	15c1	44
	15c2	44
	15c3	62
	15c4	62
	15d1	38
	15d2	62
	15d3	602
Known Obstacles	16a	132
	16b	88
	16c	72
	16d	88
Occupancy Map	20a1	14
	20a2	62
	20b1	14
	20b2	126

	20b3	14
	20c1	68
	20c2	68
	20d1	36
	20d2	44
Total		3398

7.5. Discussion

In the beginning of the research discussed in this chapter, the aim was to understand 'how to identify *"good"* test scenarios. However, it was not clear what *"good"* meant or the characteristics that defined a *"good"* test scenario in the context of ADAS and ADS. The semi-structured interview study of the verification and validation experts from the automotive industry helped to identify that one of the characteristics of a *"good"* test scenario would be one that exposes failures. Thus, the concept of Hazard Based Testing (HBT) proposed in this research focusses on exposing failures in ADAS and ADS by testing *"how a system fails"*.

Hazard identification is the first stage of HBT. There are various methods for hazard identification (discussed in section 7.2). In this thesis, it is argued that STPA, which is inspired from systems engineering, is the most effective hazard identification method for complex systems like ADAS and ADS. While the argument in support of STPA is based on literature, a possible extension of the current research would be to compare the results of various hazard identification methods (FMEA, FTA, HAZOP etc.) on the same ADAS or and ADS. It is important to acknowledge that the quality of the STPA results is dependent on the input of the system knowledge. In order to conduct the STPA of the LSAD system, the LSAD manufacturer – Aurrigo provided the system knowledge.

The proposed extension to STPA was helped by the existing structure of the Unsafe Control Actions identified in STPA. The UCA structure *("actor, control action type, control action and context")*, aided the parametrisation by highlighting the "context" element clearly.

A future research step would involve creating an algorithm to select parameter values for the scenario parameters so that the pass criteria is violated. Various white box and black box methods can be used for parameter selection. One of the promising techniques is Bayesian Optimisation (BO), which converts the parameter selection problem into an optimisation problem.

7.6. Conclusion

It is suggested in literature that to test an ADS in order to prove that it is 20% better than human driven vehicle, they need to be driven for over 11 billion miles (Kalra and Paddock, 2016a). Research discussed in this chapter found that the number of miles driven is not a meaningful metric for judging confidence in ADAS or ADS, rather the scenarios encountered by the systems are more important. It was suggested that for an ADAS or an ADS, we need to test *"how a system fails"* rather than *"how a system works"*. The concept of Hazard Based Testing (HBT) was proposed to create hazard based test scenarios. To identify hazards, Systems Theoretic Process Analysis (STPA) was used as it has been suggested in literature that it identifies more hazards as compared to other hazard identification methods like FMEA, FTA, HAZOP, ETA etc. especially for complex systems involving human-automation interaction. Grounded in systems engineering and controls engineering, STPA is a four step process which considers safety as a control problem and a breach of a control law causes an accident/loss. STPA identifies Unsafe Control Actions (UCA) and their causal reasons. This chapter proposes an extension to STPA to create test scenarios for each of the UCAs identified as a part of the STPA method. The proposed method also identifies pass criteria for the test scenarios. The proposed test scenario consists of 1) scenery 2) dynamic elements 3) additional context and 4) pass criteria. The scenery and dynamic elements are selected according the Operational Design Domain (ODD) of the vehicle. Additional context is selected from the STPA output. The proposed method was applied to a real-world case study of a Low-Speed Automated Driving (LSAD) system. The STPA analysis of a part of the Autonomous Control System of the LSAD system with nine control actions yielded 45 Unsafe Control Actions. This corresponded to the creation of 3398 test scenarios with 350 parameters associated with them.

INCREASING THE RELIABILITY OF INFORMED SAFETY[5]

Chapter 8

In the previous chapter (chapter 7), a methodology for generating test scenarios for testing automated driving systems in order to create the knowledge for *"informed safety"* was proposed. The two-pronged approach of the proposed methodology includes: requirement based testing (RBT) and hazard based testing (HBT). While the ability to generate the knowledge is an important aspect of *"informed safety"*, an equally important aspect is the ability to generate the knowledge ***reliably***, where reliability refers to the *"extent to which a framework, experiment, test, or measuring instrument yields the same results over repeated trials"* (Carmines and Zeller, 1979).

The dynamic component of *"informed safety"* in part depicts the risk associated with a particular situation. In the automotive industry, the risk associated with a hazard and the corresponding safety goal, determine the rigour in development process employed for the system. The link between the development process and the risk in the automotive industry is governed by an international standard ISO 26262 – 2018 that provides guidance to perform the automotive HARA to measure risk in terms of Automotive Safety Integrity Level (ASIL) (discussed in section 8.1). Therefore, it is essential that the ASIL ratings are reliable to ensure reliable knowledge of risk is conveyed to the drivers and also that the development process for the automated driving systems are consistent across the industry leading to establishment of true capabilities and limitations of the automated driving systems in a reliable manner. Thus, improving the reliability of drivers' informed safety. This chapter

[5] Contents of this chapter have been published in the following publication:

Khastgir, S., Birrell, S., Dhadyalla, G., Sivencrona, H., et al. (2017) 'Towards increased reliability by objectification of Hazard Analysis and Risk Assessment (HARA) of automated automotive systems', Safety Science. Elsevier Ltd, 99, pp. 166–177. doi: 10.1016/j.ssci.2017.03.024.

discusses research question 3 (RQ3: How to improve the inter- and intra-rater-reliability of the automotive HARA process).

8.1. Introduction

As discussed in chapter 6, the ISO 26262 – 2018 standard is industry's gold standard for functional safety. The standard refers to ASIL as a metric for analysing the risk. An ASIL rating comprises of three components: Severity (S), Exposure (E) and Controllability (C). However, the standard fails to provide objective guidance to evaluate the each of the three (S, E and C) ratings or identify the parameters influencing these ratings (Yu et al., 2016). The lack of objectivity leads to subjective interpretation of the standard by experts causing intra-rater and inter-rater variations leading to reduced reliability of the ratings (Ergai et al., 2016).

8.1.1. ASIL

The ISO 26262 – 2018 defines Automotive Safety Integrity Level or ASIL as *"one of four levels to specify the item's or element's necessary ISO 26262 requirements and safety measures to apply for avoiding an unreasonable risk with D representing the most stringent and A the least stringent level"*. Various ASIL levels identified by ISO 26262-2018 are QM, ASIL A, ASIL B, ASIL C, and ASIL D, where QM (quality management) denotes the lowest integrity level with no requirements to comply with ISO 26262 and ASIL D applies the most stringent requirements on product development cycle to comply with ISO 26262. The difference in requirements is also evident in Table 8.2. Based on the severity, exposure and the controllability rating, an ASIL rating is determined using the ASIL determination table specified in the ISO 26262 – 2018 Part 3 (ISO, 2018c) (Table 8.1), which shows the relation between them. The ISO 26262 standard provides ASIL dependent requirements for the development process of safety functions involving hardware and software components. The level of rigour required for higher ASIL values is considerably high as compared to a lower ASIL value. Therefore, the automotive industry is always driven towards lower ASIL values in order to keep their development costs down. This inherent bias can also sometimes lead to an inconsistency in the ASIL ratings.

Table 8.1: ASIL determination table (adapted from ISO 26262 – 2018: Part 3 (ISO, 2018c)*)*

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

The difference in the requirements for development processes to be followed for various ASIL levels is mentioned in the standard via many tables. Table 8.2 illustrates the increased rigour required in the methods for software unit testing as the ASIL level increases. For an ASIL C and ASIL D system, back-to-back comparison test between model and code is highly recommended as per the standard which adds considerable cost to the product development cycle.

Table 8.2: Methods for verification of software integration (adapted from ISO 26262-2018: Part 6 (ISO, 2018d)*)*

	Method	ASIL A	ASIL B	ASIL C	ASIL D
1a	Requirements-based test	++	++	++	++
1b	Interface test	++	++	++	++
1c	Fault injection test	+	+	++	++
1d	Resource usage evaluation	++	++	++	++
1e	Back-to-back comparison test between model and code, if applicable	+	+	++	++
1f	….				
++ : highly recommended; + : recommended; o : no recommendation for or against					

8.1.2. Severity

The ISO 26262 – 2018 defines "severity" as *"estimate of the extent of harm to one or more individuals that can occur in a potentially hazardous situation"*, for the driver or the passengers of the vehicle or other vulnerable road users like cyclists, pedestrians in the vicinity of the vehicle. The standard refers to the Abbreviated Injury Scale (AIS) (Baker et al., 1974) as one of the methods for calculating the severity rating. The standard defines four classes for severity: 1) S0 (no injuries) 2) S1 (Light and moderate injuries) 3) S2 (severe and life threatening injuries) 4) S3 (life-threatening injuries, fatal injuries)

8.1.3. Exposure

The ISO 26262 – 2018 defines "exposure" as *"state of being in an operational situation that can be hazardous if coincident with the failure mode under analysis"*. The standard defines five classes for exposure: 1) E0 incredible 2) E1 (very low probability: Occurs less often than once a year for the great majority of drivers) 3) E2 (low probability: Occurs a few times a year for the great majority of drivers) 4) E3 (medium probability: Occurs once a month or more often for an average driver) 5) E4 (high probability: occurs during almost every drive on average).

8.1.4. Controllability

The ISO 26262-2018 (ISO, 2018c) standard states that *"The evaluation of the controllability is an estimate of the probability that someone is able to gain sufficient control of the hazardous event, such that they are able to avoid the specific harm."*.

While the standard classifies controllability into four classes: 1) C0 (Controllable in general) 2) C1 (simply controllable: 99 % or more of all drivers or other traffic participants are usually able to avoid harm) 3) C2 (normally controllable: 90 % or more of all drivers or other traffic participants are usually able to avoid harm) 4) C3 (difficult to control or uncontrollable: less than 90 % of all drivers or other traffic participants are usually able, or barely able, to avoid harm), it fails to elaborate on the criteria for the classification and defining the levels in a more objective manner. This introduces a degree of vagueness and subjectivity to the classification. To give a rating for controllability, the experts needs to understand how a driver/operator would react to a hazard caused by a failure for any given situation to have a valid rating. As discussed in chapter three, such an analysis will be based on the expert's mental model and background knowledge leading to inter-rater variation, as the assumptions and mental models may differ significantly between experts. The two distinct short-comings of the current ISO 26262-2018 standard are guided by the subjective nature of the experts' mental models leading to unreliable ratings and the inability to identify all hazards (including the black swan events). Additionally, controllability argument changes when an autonomous system is considered as the driver is no longer a fall-back option.

8.1.5. Reliability through objectivity

According to Cambridge English Dictionary ("Cambridge English Dictionary," 2017), "objectivity" is defined as "*the state or quality of being objective and fair*", where "objective" is defined as "*based on real facts and not influenced by personal beliefs or feelings*".

In order to prevent the influence of personal beliefs and mental models of experts leading to varied and unreliable HARA ratings and to answer the research question three, the author proposes the creation of a rule-set to introduce objectivity in the process. Objectivity could potentially be a tool to help provide consistency and convergence of HARA ratings, thus providing increased reliability.

8.2. Creating Rule-set

To answer research question three, the author proposes objectivising severity and controllability ratings by introducing a rule-set for both the ratings. One of the first steps was to run a pilot study (discussed in Appendix A4) to evaluate the existence of reliability issues in HARA ratings in the automotive domain. This study was one of the first steps towards achieving reliable ratings through an objective decision making process for HARA.

The initial rule-set (used in the pilot study) didn't provide rules for exposure rating. The author in his analysis of hazards with a different set of experts had come to a conclusion that the exposure rating for the given hazardous events and the given system (discussed in section 8.3.3.2 and section 8.3.3.1) will most certainly be E4 (highly probable). In order to objectify the HARA process, severity and controllability ratings' rule-set were parametrized in terms of factors (parameters) identified by the author. While various hazards and hazardous events were identified, various parameters were used to classify a hazardous event. These included acceleration value, velocity etc. The first set of parameters were identified from this set. In addition, existing literature was reviewed for factors influencing severity and controllability (Baker et al., 1974; Ellims and Monkhouse, 2012; Green, 2000; Lortie and Rizzo, 1998; Monkhouse et al., 2015; Summala, 2000; Verma and Goertz, 2016). The parametrization of the HARA components should help meet the R1, R2 and R3 reliability criteria defined by Aven and Heide (2009), by objectivising the decision making process involved in HARA ratings. Figure 8.1 depicts the process of development of the initial rule-set, along with stakeholder roles at each step. Feedback on the rule-set was received from independent functional safety experts.

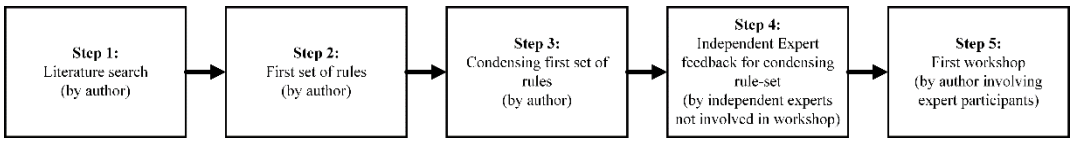

Figure 8.1: Process of developing initial rule-set with role description for each step

8.2.1. Severity rating rule-set

The severity parameters were mainly influenced by impact energy, characteristics of impact and the environment (Johansson and Nilsson, 2016a). Therefore, the parameters identified for severity rating were: 1) vehicle velocity 2) oncoming object velocity 3) type of obstacle 4) type of impact (side, head-on etc.) 5) gradient of slope 6) magnitude of delta torque (difference between required and provided torque) 7) maximum acceleration/deceleration 8) mass of vehicle. However, the severity rule-set depicted in Table 8.3 is a condensed version of the initial rule-set. A condensed version of the rule-set (prepared by the author) was used due to logistical reasons of conducting the validation of the rule-set. The condensed version of the rule-set was prepared by deleting some of the secondary parameters like type of collision (head-on, side, rear), gradient of slope, country/city for which the hazard was described for etc.

Table 8.3: Initial Severity rule-set for US workshop (workshop 1)

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Severity Rating
Pedestrian	< 11 km/h	< 2 km/h	S0
		< 6 km/h	S1
		< 12km/h	S1
	11 - 16 km/h	< 2 km/h	S1
		< 6 km/h	S2
		< 12km/h	S2
	> 16 km/h	< 2 km/h	S2
		< 6 km/h	S3
		< 12km/h	S3

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Severity Rating
Infrastructure	< 11 km/h	0 km/h	S0
		0 km/h	
		0 km/h	
	11 - 16 km/h	0 km/h	S1
		0 km/h	
		0 km/h	
	> 16 km/h	0 km/h	S2
		0 km/h	
		0 km/h	

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Severity Rating
Vehicle	< 11 km/h	< 10 km/h	S0
		< 20km/h	S1
		> 20 km/h	S2
	11 - 16 km/h	< 10 km/h	S1
		< 20km/h	S1
		> 20 km/h	S2
	> 16 km/h	< 10 km/h	S1
		< 20km/h	S2
		> 20 km/h	S3

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Severity Rating
Cyclist	< 11 km/h	< 8 km/h	S0
		< 14km/h	S1
		< 20km/h	S2
	11 - 16 km/h	< 8 km/h	S1
		< 14km/h	S2
		< 20km/h	S2
	> 16 km/h	< 8 km/h	S2
		< 14km/h	S2
		< 20km/h	S3

8.2.2. Controllability rating rule-set

One of the major factors influencing controllability of the hazardous situation is the speed of the vehicle. In order to evaluate the effect of speed on controllability of a vehicle, a driving simulator experiment was conducted. Details of the experiment are discussed in Appendix A4. The findings of the study two (controllability study) (discussed in Appendix A4) were incorporated in the controllability rating rule-set. The controllability parameters were mainly influenced by the vehicle's ability to change trajectory and the environment affecting vehicle's ability to make this change (McGehee et al., 2000; Rosén et al., 2011; Schaap et al., 2008; Young and Stanton, 2007). The parameters identified for controllability were: 1) vehicle velocity 2) time-to-collision (TTC) 3) distance to obstacle 3) maximum acceleration/deceleration 4) availability of safe area 5) road friction 6) gradient of slope. Time-to-collision (TTC) is defined as *"the time taken by the trailing vehicle to crash into the front vehicle, if the vehicles continue in the same path without adjusting their speeds"* (Chin and Quek, 1997). Similar to the severity rule-set, a condensed version of the controllability rule-set was used in the pilot study due to logistical reasons and is depicted in Table 8.5. The condensed version was prepared on similar basis as the severity rule-set (in addition to the results from controllability study (study two: discussed in Appendix A4)), and is illustrated in Table 8.4.

Table 8.4: Initial Controllability rule-set for US workshop (workshop 1)

Emergency Deceleration Value	Distance to Obstacle	TTC	Vehicle velocity	Controllability Rating
0.4g - 0.8g	< 6 m	< 1.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C0
			> 16 km/h	C2
	> 6 m	< 1.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C2
		1.0 - 2.0 sec	< 11 km/h	C0
			11 - 16 km/h	C0
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C0
			> 16 km/h	C1

Emergency Deceleration Value	Distance to Obstacle	TTC	Vehicle velocity	Controllability Rating
< 0.4g	< 6 m	< 1.0 sec	< 11 km/h	C3
			11 - 16 km/h	C2
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C2
			> 16 km/h	C3
		> 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C3
	> 6 m	< 1.0 sec	< 11 km/h	C3
			11 - 16 km/h	C2
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C3
		> 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C2

8.3. Method

In order to answer the research question three (identified in chapter six) (*"How to improve the inter- and intra-rater-reliability of the automotive HARA process?"*), the author created rule-sets for severity, controllability ratings and subsequently for exposure ratings too. A hybrid approach was adopted for creating the rule-set. For creating the controllability ratings, a driving simulator study was conducted to evaluate the effect of vehicle speed on controllability. Additional parameters were identified based on literature review. The creation of the initial rule-set is discussed in section 8.2.

Once the initial rule-set was created, to test the hypothesis that a rule-set could increase the objectivity of the automotive HARA process, potentially leading to convergence in ratings and thus improving the reliability of the automotive HARA, a series of workshops involving international functional safety experts were conducted (Figure 8.2). The workshop studies were conducted in US, Sweden, Germany and the UK.

The first workshop study (pilot study) was conducted in United States at a conference to which the author was invited as an invited speaker. The first workshop was essentially a scoping workshop to understand two aspects of the study:

- Do industry experts share the opinion on the reliability challenges of the automotive HARA?
- Do industry experts believe that providing rules for automotive HARA can be beneficial to achieve increased reliability of the HARA process?

While the industry experts appreciated the reliability challenges of the automotive HARA and the use of rules to overcome the challenges, it was mentioned that the rules needed to be more elaborate to avoid their misinterpretation. The detailed feedback and the results are discussed in section 8.4.3. Based on the results and participant feedback from the first workshop, the second workshop was conducted in Sweden. While better convergence in HARA ratings was achieved in the Sweden workshop, controllability ratings still showed variation which was also mentioned in the feedback from the experts. This feedback helped to further refine the workshop design and the rules. The penultimate workshop (workshop 3) was held in Germany on the side-lines of a conference where the author was invited as an invited speaker. After workshop 3, the rule-set was re-calibrated based on the feedback from the workshop 3 and independent functional safety experts who had not taken part in any of the workshops. With the final rule-set, the last workshop was conducted with a modified workshop design in the UK at WMG, University of Warwick's premises with functional safety experts from the UK's automotive industry.

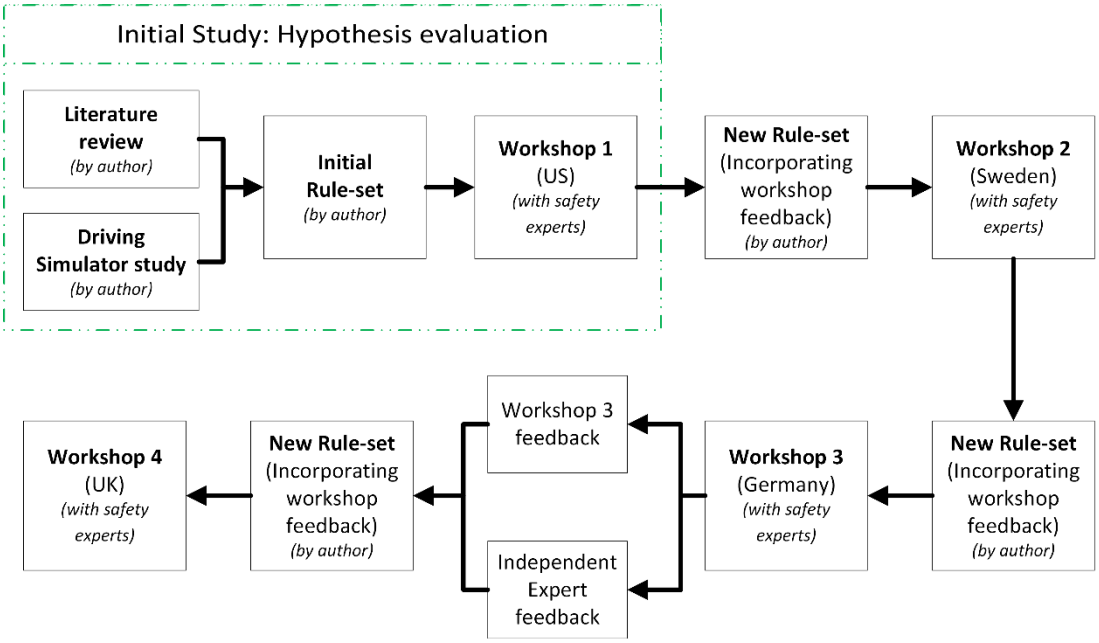

Figure 8.2: Timeline for HARA workshops and rules development

8.3.1. Ethical Approval

Ethical approval for the series of workshops was secured from the University of Warwick's Biomedical & Scientific Research Ethics Committee (BSREC) (BSREC Ethical Application ID: REGO-2016-1845). All data gathered from the workshop was treated in a confidential manner, in accordance with the University of Warwick's Data Protection Policy. Informed consent was obtained from all participants.

8.3.2. Participants

The number of participants in each of the workshops varied from 10 – 17. However, in each of the workshops, participants were divided into different groups. Depending on the number of participants taking part in the workshop, participants were arranged in either three groups (US and UK workshop), four groups (Germany workshop) or five groups (Sweden workshop).

8.3.3. Workshop structure

The workshop was modelled on the World Café method (Fouche and Light, 2011). Putting participants in a world café style workshop keeps them engaged in the given task. Additionally, it also allows participants to mix with each other which was an important reason for the choice (to evaluate inter-rater reliability variation).

Workshop one (US workshop), two (Sweden workshop) and three (Germany workshop) had four rounds each, while workshop four (UK workshop) had two rounds only. Having already established the inter-rater variation caused by mixing participants in groups, the focus of workshop four was only on the impact of the rule-set and its potential to increase the reliability of the automotive HARA process, which meant that two of the rounds could be removed. The detailed structure for individual workshops is discussed in sections 8.4.2, Appendix A4, 8.6.2 and 8.7.2.

8.3.3.1. System definition

Participants were asked to perform a HARA for a given hazard and hazardous events for a Low Speed Autonomous Driving System (LSAD), i.e. a pod. Participants were asked to make the assumption that the current ISO 26262 Part 3, which is an automotive functional safety standard for passenger vehicles is applicable for the LSAD (pod).

8.3.3.2. Hazard and hazardous event definition

According the international standard ISO 26262 – 2018, hazard is defined as *"potential source of harm caused by malfunctioning behaviour of the item"*. In addition ISO 26262 – 2018 defines hazardous event as *"combination of a hazard and an operational situation"*, where operational situation is defined as *"scenario that can occur during a vehicle's life"*.

For the first workshop, the hazard provided was identified after conducting in-depth hazard analysis for a low-speed autonomous vehicle and a qualitative analysis was carried out on the explanation for the analysis. The in-depth hazard analysis was conducted by independent functional safety experts involved in the UK Autodrive[6] project. The hazard and the hazardous events definition for the pod was a result of this HARA. Various functions like Torque management, braking and route planning could cause a given hazard. All functions causing the hazard (provided to the experts in the workshops) were related to vehicle's movement. However, the hazard definition was modified to incorporate the feedback from the experts from each of the workshops.

[6] http://www.ukautodrive.com/

8.3.3.3. Qualitative feedback

At the end of workshop rounds, each group was asked to provide feedback on the workshop by answering two questions:

- (During the workshop) Have you experienced variation in hazard analysis discussions based on the group of people involved in the discussion?
- Do you think by having rules by parametrizing hazard analysis, we can have a more objective approach?

8.4. Workshop 1 (USA: Initial Scoping workshop)

8.4.1. Participants

Twelve participants were involved in the workshop, who had experience in automotive functional safety assessments. Eight out of the 12 participants identified themselves as automotive functional safety specialists and had taken part in international ISO 26262 functional safety technical committee discussions. The remaining four participants identified themselves as development/systems engineers applying automotive functional safety principles in their function development process. Participants had diverse backgrounds representing different levels of supply chain across the automotive supply chain. Two participants were from OEM (original equipment manufacturer), seven were from Tier One suppliers, two were from Tier Two suppliers and remaining one participant was from academia/research organization background. All participants were from North America and Europe.

8.4.2. Workshop 1 structure

In workshop 1, participants were grouped into three groups of four participants each. The workshop consisted of an introduction which was followed by four rounds of 25 minutes each. Each group was provided with two different hazardous events and were asked to rate the two given hazardous events. The same hazardous events were given in each of the four rounds. Figure 8.3 shows the workshop 1 structure. Participants were informed about the system (section 8.3.3.1 and section 8.4.2.2) for which they were being asked to perform the HARA during the introduction stage.

Before starting the rounds of discussion for HARA, each group (assigned a table) was asked to nominate one participant as the moderator for the group. In round one, each group was

supposed to discuss and come to a consensus for each of the two hazardous events, on a rating for Severity (S), Exposure (E) and Controllability (C) and subsequently for ASIL. After round one, the members of the groups were shuffled, but the moderator for each group remained same. The shuffling was done in a way that the table had at least two new participants as compared to the previous round. In round two, the new groups were asked to discuss and give ratings for S, E and C. After round two, participants were provided with a rule-set by the author for conducting HARA. The participants were instructed how to use the rule-set. Participants were instructed not to question the rules for their validity. However, they were given the freedom to interpret the rules as per their understanding. In round three, participants used the provided rule-set for HARA to complete the task of S, E and C ratings for the two hazardous events. The groups were same in round two and round three. After round three, the groups were again shuffled, but the moderators for the groups remained the same. In round four, the new groups were again tasked to use the rule-set for HARA (provided to them) to rate the two hazards for S, E and C. The mixing of groups after round one and round three helps address the research question of inter-rater variability (with and without the rule-set). Moderators were asked to provide a brief explanation of the discussion in each round and the reasoning behind the rating for each of the parameters (S, E and C).

This provided a possibility to perform both quantitative and qualitative analysis on the gathered data which includes the ratings in each round (quantitative) and the moderators' explanation in each round (qualitative).

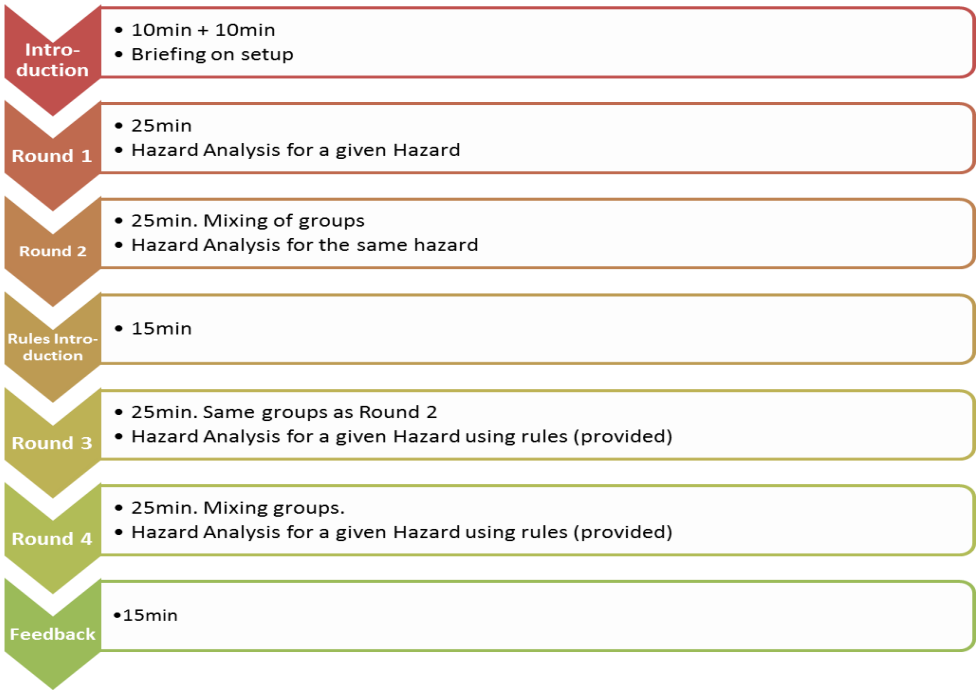

Figure 8.3: Workshop 1 structure

8.4.2.1. System definition

Participants were asked to perform a HARA for the provided hazard and hazardous events for a Low Speed Automated Driving Vehicle (LSAD). The system features presented to the participants were:

- Fully Autonomous in dedicated ODD (SAE Level 4)
- Connected vehicle with Vehicle-to-Infrastructure (V2I) capability
- Emergency stop button. No trained safety driver
- No steering wheel or pedals
- Top speed of 25 km per hour

Participants were asked to make the assumption that the current ISO 26262-2011 (in 2016) Part 3 (2011 was the latest version of the standard at the time of the workshop), which is an automotive functional safety standard for passenger vehicles is applicable for LSAD system. Participants were advised to use the ASIL determination table, which was provided to them during the workshop from the mentioned standard.

8.4.2.2. Hazard definition

The hazard provided to the participants was *"Collision (of pod) with static or dynamic obstacle due to stopping or accelerating to a vulnerable position"*. Based on the hazard, participants were provided two hazardous events and were asked to discuss the HARA for the two given events to give S, E and C ratings. The two hazardous events provided were:

- Pod travels into pedestrian / cyclist
- Pod does unintended braking

8.4.3. Results

8.4.3.1. Quantitative Results

Each group was asked to provide a rating for Severity, Exposure and Controllability for the two hazardous events for each round of their discussion. Figure 8.4 shows the ASIL ratings provided by the individual groups in different rounds. Different rounds have been plotted on the x-axis and the ASIL ratings have been plotted on the y-axis. Rules for HARA were provided only in round three and round four. In the first round, (when no rules were provided to the participants), each group came up with a different ASIL rating with significant differences. The difference between the groups were of the order of two for group 1 and group 3 (ASIL A and ASIL C for first hazardous event) and group two and group

three (QM and ASIL B for second hazardous event). The difference with the other group was of the order of one. Round two proved to have some convergence in the ratings, however there were still significant differences in the ASIL ratings. For hazardous event 1, two groups converged to an ASIL rating of ASIL C, while the third group differed significantly with an ASIL rating of QM which means the difference was of the order three. For hazardous event 2, while two of the groups converged to an ASIL A rating, the third group gave a QM rating which meant a difference of the order of one. It is interesting to observe that the groups giving QM rating to hazard 1 and hazard 2 were different.

The variation in the ASIL ratings provided by the groups in round one and round two, illustrates the low reliability (inter-rater) of the current automotive hazard analysis method, even when done by experts in the industry. While every group was provided with the same hazardous events to rate, each of them had a different justification for the ASIL rating provided by them. The difference demonstrates the inter-rater variability in automotive HARA due to presence of subjectivity which is caused by the experts' mental models. This makes the HARA process unreliable as per the R2 and R3 criteria of reliability mentioned by Aven and Heide (2009). The variation in the HARA ratings will be discussed in more detail in the qualitative analysis section (section 8.4.3.2).

Before round 3, rules for HARA were introduced to the participants and they were asked to use the rules to perform the HARA. It was expected that the introduction of the rule-set would introduce objectivity in the HARA process and potentially lead to a convergence in the ASIL ratings from the three groups of experts. However, the results (as depicted in Figure 8.4), illustrate the opposite. In round three, for both hazardous event 1 and hazardous event 2, the three groups provided three different ASIL ratings with a maximum difference of order two and the minimum difference of order one. This was contrary to the expectation of the author. However, the qualitative analysis of the round three results (section 8.4.3.2) provide a deeper insight on the cause of the variation. Round four provided an interesting set of results for hazardous event 1 and hazardous event 2, with convergence in ratings achieved for hazardous event 2.

The ASIL ratings for hazardous event 1 between rounds 1-2 and rounds 3-4, show a visual decrease in variation (Figure 8.4), indicating shift towards convergence, potentially due to the introduction of rule-set. In an ideal situation, for a fully reliable HARA, the variation in ratings should be zero. While ASIL ratings for hazardous event 1 provided by different groups varied significantly (with a maximum variation of order 2 and a minimum variation of order 1), ASIL ratings for hazardous event 2 converged for all groups at ASIL A. At a higher level, it might seem that the convergence of the ASIL rating for hazardous event 2 is

a result of the introduction of the rule-set by the author. However, a more granular analysis of the components of ASIL provides a different view. As discussed in section 8.1.1, an ASIL rating is comprised of a severity rating (S), exposure rating (E) and a controllability rating (C). Figure 8.5 – 8.7 depict the severity, exposure and the controllability ratings respectively for hazard 1 and hazard 2.

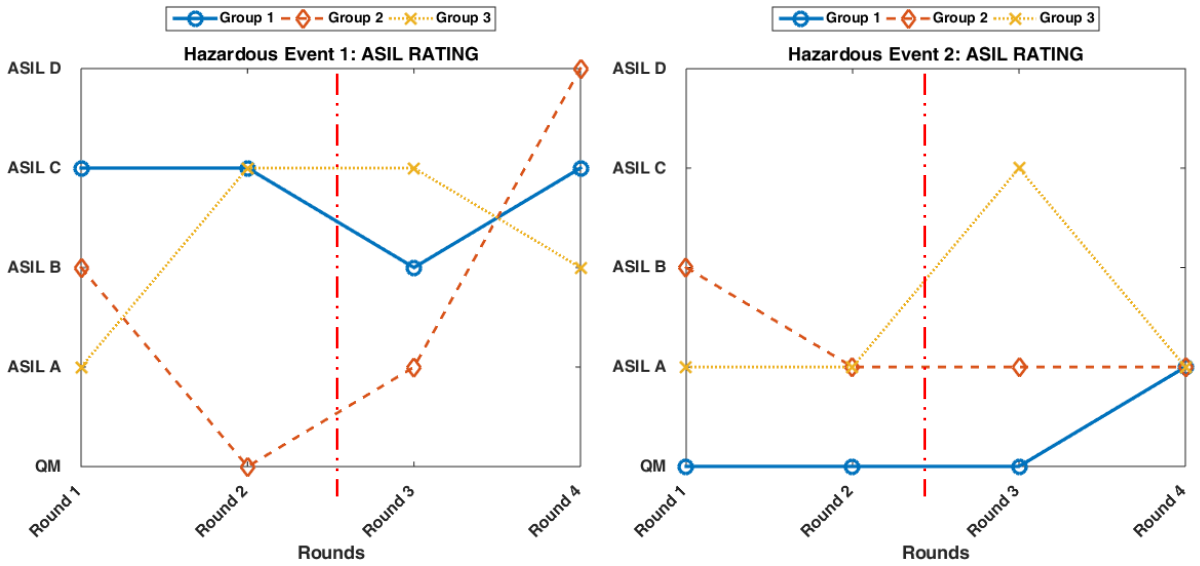

Figure 8.4: ASIL ratings for hazard 1 and hazard 2 given by experts in different rounds (Round 1 and Round 2: without rule-set); (Round 3 and Round 4: with rule-set)

Severity

In round 1, while two groups agreed on the severity rating, the third group provided a rating with a difference of order two for hazardous event 1 (Figure 8.5). In round two, all the groups converged in their severity rating at S3 for hazardous event 1. With the introduction of rules in round three, while two of the groups converged in their severity rating at S2 (which was different from their round two ratings), the third group gave a rating (S3) which differed in the rating of the other two groups by the order of one. In round four, after the groups were mixed, a similar spread was found with two groups agreeing in their severity rating at S2, while the third group gave a rating of S3. The group giving a diverging rating to the others was different in round three and round four. For hazardous event 2, two groups converged completely across all the rounds. However, the third group showed significant variation across the rounds. In round one, the severity rating of the third group was in agreement with the other groups at S1. However, in round two, the group gave a rating of

S2. With introduction of rules, the group gave a severity rating of S3 and S2 in round three and round four respectively.

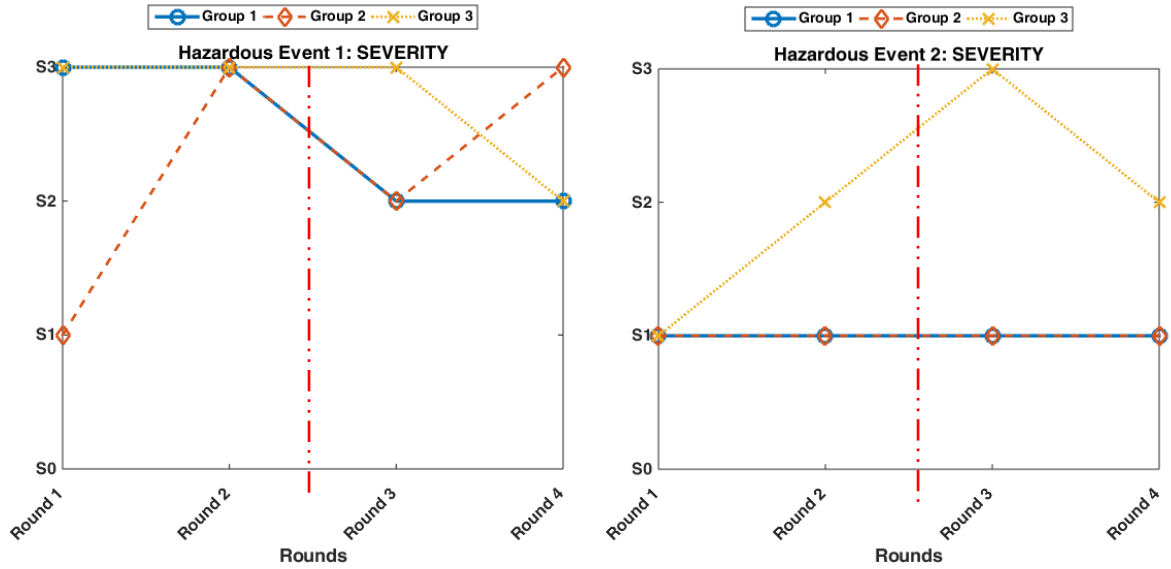

Figure 8.5: Severity ratings for hazard 1 and hazard 2 given by experts (in different rounds) Round 1 and Round 2: without rule-set; Round 3 and Round 4: with rule-set

Exposure

In the workshop experiment, no rules were provided for exposure rating. While this was due to the author's understanding of exposure rating being almost certainly being constant, the experiment was also designed to see if there was any intra-rater variability, i.e., variation in the same group of people with experience. In case any intra-rater variance was present, this would be seen in the ratings of round 2 and round 3, as the groups in the two rounds were identical. While there was no evidence of intra-rater variability in the exposure ratings, a significant degree of inter-rater variability existed among the different groups across various rounds (Figure 8.6). Contrary to the author's hypothesis, the variation in exposure ratings was high, as compared to the severity and the controllability ratings for hazardous event 1. While the same was true for rounds 1-2 for hazardous event 2, rounds 3-4 for hazardous event 2 showed the least variation for exposure rating.

Controllability

Controllability ratings for hazardous event 1 showed a similar variation as that of the severity ratings. However, the variation for controllability ratings rose for both the hazardous events, with the introduction of the rules. This could potentially be due to the interpretation of the rules provided to the participants.

Ideally, the introduction of the rule-set for HARA should have led to zero variation in the severity, exposure, controllability and ASIL ratings. While the reduction was observed in some of the ratings (Figure 8.7), it is important to analyse the results qualitatively (section 8.4.3.2) to explain the deviation.

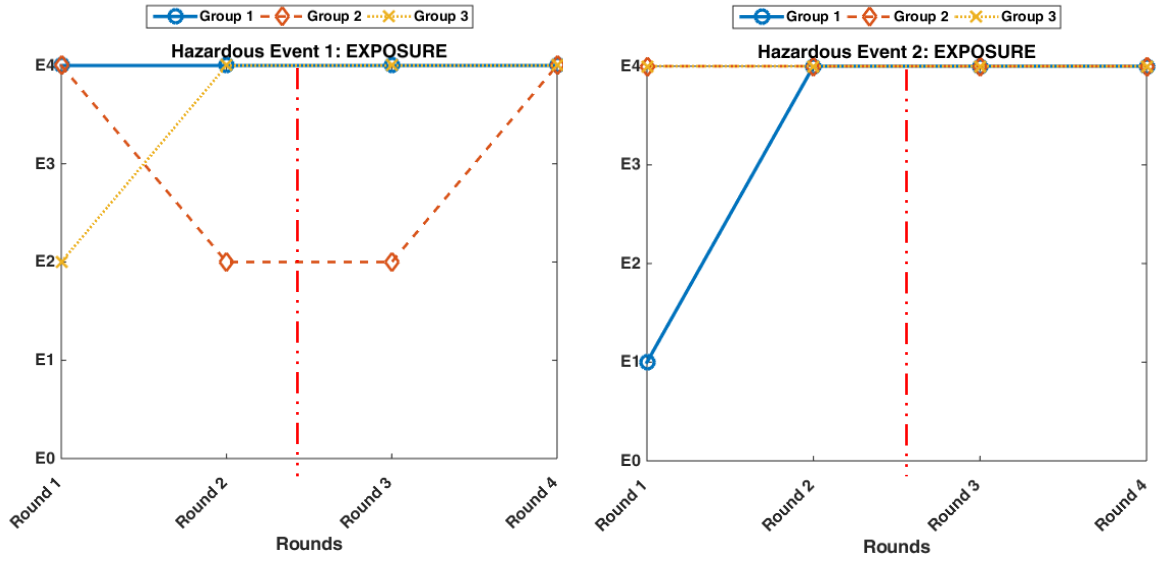

Figure 8.6: Exposure ratings for hazard 1 and hazard 2 given by experts in different rounds. Round 1 and Round 2: without rule-set; Round 3 and Round 4: with rule-set

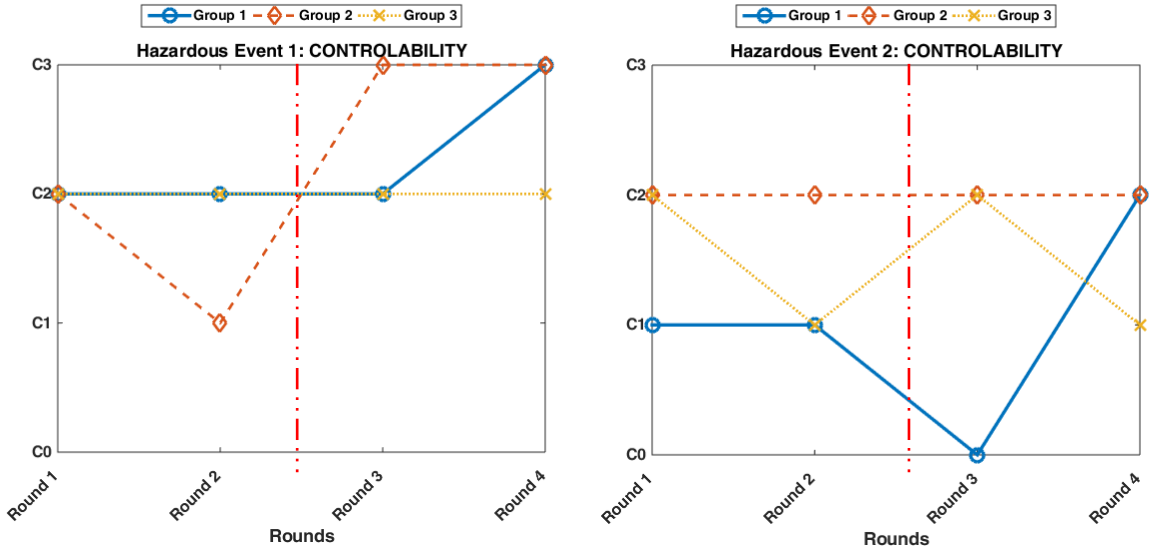

Figure 8.7: Controllability ratings for hazard 1 and hazard 2 given by experts in different rounds). Round 1 and Round 2: without rule-set; Round 3 and Round 4: with rule-set

8.4.3.2. Qualitative Results

Each of the three groups were asked to provide answers to the questions mentioned in section 8.3.3.3 about their experience of HARA in the different rounds of the workshop. While answering the first question about experiencing variation in hazard analysis discussions, all three groups mentioned that they had experienced variation in HARA discussions in different rounds. All three groups concurred that the source of variation was the different perspectives presented by different individuals present in the group. However, the reasons for varying perspectives differed between the groups. One of the groups mentioned that the HARA is dependent on persons' experience and their previous training/understanding of the rating procedure in HARA. This coincides with the literature discussed earlier (in chapter 6) about the background knowledge of the experts being one of the reasons for subjectivity (Aven and Zio, 2014). Another group mentioned that experts from different cultures, perceived *"severity"* and *"exposure"* ratings differently and there is a need to provide context regarding the environment for which the product is being made. Although, limited literature exists to support the cultural factor as a source of subjectivity in HARA, recent studies in other domains like occupational health and safety (OHS) have indicated this trend (Aven and Zio, 2014; Tchiehe and Gauthier, 2017). Having participants from North America and different European countries was beneficial in observing this trend.

Two out of the three groups agreed in their response to the second question by saying that the introduction of rules by parametrizing HARA made the process more objective. While the third group disagreed with the statement, but qualified their response by mentioning that the rules, the parameters and their relationship were open to subjective interpretation. The other two groups mentioned that the rules needed to be re-calibrated in certain areas (like introducing context for the rules) and more examples and instructions need to be provided before using the rules. This is further established by the fact that each of the groups in round three and four (while using the rules for HARA) made different initial assumptions about the system and the hazard due to which they came to a different severity and controllability rating. This emphasizes the importance of the initial assumptions made by the experts performing the HARA and was also highlighted by one of the groups in their feedback. Providing context to the rule-set could potentially help to remove the subjective nature of the initial assumptions and will be introduced in future workshop studies.

8.4.4. Learnings from workshop 1: A discussion

One of the aspects highlighted in the qualitative analysis of the feedback was on the need for a few example cases and training to use the provided rule-set. This would potentially aid the experts' understanding on how to use the rule-set provided for performing HARA. In order

to overcome the challenge due to unclear understanding of the process, based on the feedback from workshop 1, the author extended the rule-set introduction time during future workshop (section 8.5 – 8.7) and also incorporated a few example cases. The participants also suggested the inclusion of some of the assumptions about the crashworthiness of the vehicle and the operational design domain of the vehicle.

Since, contrary to the author's hypothesis, it was found that the exposure ratings were also subject to high degree of variation, an additional rule-set for exposure ratings was introduced in the next workshops (discussed in sections 8.5, 8.6 and 8.7). At the time, it was believed that an exposure rule-set along with the context definition should potentially be able to bring convergence in the exposure ratings and hence ASIL ratings too.

The hazard and the hazardous events chosen for the workshop study were a small part of a larger collection of hazards and hazardous events. The full collection was created as a result of a safety analysis of the low-speed autonomous vehicle. While the independent group of experts who performed the safety analysis had full information about the system and the hazards, the expert participants in the workshop study had limited information about given hazard. In some of the qualitative feedback, participants mentioned the need for more information. However, it was also noticed from the discussion notes of the expert panels that they found it hard to implement the classification method. In order to mitigate such instances, hazard and hazardous events with more information about the situation and context were provided in future workshops

8.5. Workshop 2 (Sweden)

After the initial workshop in USA, the next workshop was conducted in Sweden. The workshop was attended by seventeen participants. All of the participants had prior experience in functional safety assessment. The workshop had a world café structure (similar to workshop one: section 8.4). However as the number of participants were larger than the first workshop, experts were divided into five groups. Detailed study methodology and results are discussed in Appendix A5.

8.5.1. Learnings from workshop 2: A discussion

Most of the participants in workshop 2 suggested incorporating an additional parameter of *"type of collision (head-on, rear, side-on)"* in the severity rule-set. Moreover, some suggested that the controllability rule-set was over dimensioned and *"on-coming object velocity"* could be removed as a parameter. These comments suggested that experts considered the validity of the rule-set of as an important part of the exercise, even though it

was clearly stated to participants that they should assume that the provided rule-set is valid. In a repeat from the first workshop, one of the participants commented that a user guidebook on how to use the rule-set would be helpful. For the next workshop, an even more elaborate introduction on how to use the rule-set was provided.

Even though the hazard and hazardous event definitions were reviewed and agreed upon by independent functional safety experts, participants in the workshop raised questions about the definition and need for more specificity in the hazardous event definition. According to some of the participants, the lack of specificity (e.g. operational conditions) led to subjective assumptions on those parameters which decreased the reliability of the HARA process even with the use of the rule-set.

8.6. Workshop 3 (Germany)

8.6.1. Participants

Twelve participants (including moderators) with experience in system safety took part in workshop three. The participants were from diverse backgrounds with two of participants identifying their affiliation to an OEM, five participants to a Tier 1 supplier, one to a Tier 2 supplier and three to academia/research organization. The average work experience of the participants (including moderators) was more than 16 years. Eight out of the 12 participants identified themselves as functional safety specialists, one of them identified as Quality Assurance Manager and two others identified themselves as system engineers. Participants were from Europe, United States and Australia.

8.6.2. Workshop structure

The 12 participants were grouped into four groups (each group assigned a table) with each group having three participants. Each of the groups had a moderator who was the table owner and was decided prior to the start of the workshop. Similar to workshops one and two, workshop three also had four rounds. Each round was of 25 minutes. As in workshop one and two, each group was provided with two different hazardous events and was asked to rate the two given hazardous events. The same hazardous events were given in each of the four rounds. In each round, the participants were asked to provide Severity (S), Exposure (E) and controllability (C) ratings for the two hazardous events. The workshop structure and participant shuffling was similar to workshop 1 (Figure 8.3), with participants being shuffled at the end of round 1 and round 3. Rules for HARA were introduced at the end of round 2 and participants were asked to use the rules for assigning the S, E and C ratings in rounds 3 and 4.

8.6.2.1. Additional System definition

In addition, to the system definition for the SAE Level 4 Low-Speed Automated Driving system (LSAD) described in section 8.3.3.1, participants were also provided with the following description:

- Vehicle crashworthy up to top speed
- Driving domain: public roads in inner city in UK

8.6.2.2. Hazard definition

The hazard provided to the participants was *"Unintended acceleration to a vulnerable position (in Germany)"*. Based on the hazard, participants were provided two hazardous events and were asked to discuss the HARA for the two given events to give Severity, Exposure and Controllability ratings. The two hazardous events provided to the participants were:

- Pod collides with pedestrian (child) in a side impact.
- Pod is in roadway when vehicle approaches

8.6.3. Rule-set

8.6.3.1. Severity rule-set

Based on the feedback from the Sweden workshop, *"type of impact"* was added as one of the parameters for severity rating. Table 8.5, Table 8.6 and Table 8.7 illustrate the modified severity rule-set.

Table 8.5: Severity rule-set (part 1) for Germany workshop (workshop 3)

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Type of Impact	Severity Rating
Pedestrian (Adult)	< 11 km/h	< 2 km/h	Head-on	S0
			Rear	S0
			Side	S0
		< 6 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 12km/h	Head-on	S1
			Rear	S1
			Side	S1
	11 - 16 km/h	< 2 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 6 km/h	Head-on	S2
			Rear	S1
			Side	S1
		< 12km/h	Head-on	S2
			Rear	S1
			Side	S2
	> 16 km/h	< 2 km/h	Head-on	S2
			Rear	S1
			Side	S2
		< 6 km/h	Head-on	S3
			Rear	S2
			Side	S2
		< 12km/h	Head-on	S3
			Rear	S2
			Side	S3
Type of Obstacle	**Vehicle Velocity**	**Oncoming Obj. Velocity**	**Type of Impact**	**Severity Rating**
Pedestrian (Child)	< 11 km/h	< 2 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 6 km/h	Head-on	S1
			Rear	S1
			Side	S1
		< 12km/h	Head-on	S1
			Rear	S1
			Side	S1
	11 - 16 km/h	< 2 km/h	Head-on	S2
			Rear	S1
			Side	S2
		< 6 km/h	Head-on	S2
			Rear	S2
			Side	S2
		< 12km/h	Head-on	S3
			Rear	S2
			Side	S3
	> 16 km/h	< 2 km/h	Head-on	S3
			Rear	S2
			Side	S2
		< 6 km/h	Head-on	S3
			Rear	S2
			Side	S3
		< 12km/h	Head-on	S3
			Rear	S3

Table 8.6: Severity rule-set (part 2) for Germany workshop (workshop 3)

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Type of Impact	Severity Rating
Infra-structure	< 11 km/h	0 km/h	Head-on	S0
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
	11 - 16 km/h	0 km/h	Head-on	S1
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
	> 16 km/h	0 km/h	Head-on	S2
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
Type of Obstacle	**Vehicle Velocity**	**Oncoming Obj. Velocity**	**Type of Impact**	**Severity Rating**
Cyclist	< 11 km/h	< 8 km/h	Head-on	S0
			Rear	S0
			Side	S0
		< 14 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 20km/h	Head-on	S2
			Rear	S1
			Side	S2
	11 - 16 km/h	< 8 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 14 km/h	Head-on	S2
			Rear	S1
			Side	S2
		< 20km/h	Head-on	S2
			Rear	S2
			Side	S2
	> 16 km/h	< 8 km/h	Head-on	S2
			Rear	S1
			Side	S2
		< 14 km/h	Head-on	S2
			Rear	S2
			Side	S2
		< 20km/h	Head-on	S3
			Rear	S2
			Side	S3

Table 8.7: Severity rule-set (part 3) for Germany workshop (workshop 3)

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Type of Impact	Severity Rating
Vehicle	< 11 km/h	< 10 km/h	Head-on	S0
			Rear	S0
			Side	S0
		< 20 km/h	Head-on	S1
			Rear	S0
			Side	S1
		> 20km/h	Head-on	S2
			Rear	S1
			Side	S2
	11 - 16 km/h	< 10 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 20 km/h	Head-on	S1
			Rear	S1
			Side	S1
		> 20km/h	Head-on	S2
			Rear	S2
			Side	S2
	> 16 km/h	< 10 km/h	Head-on	S1
			Rear	S1
			Side	S1
		< 20 km/h	Head-on	S2
			Rear	S2
			Side	S2
		> 20km/h	Head-on	S3
			Rear	S2
			Side	S3

8.6.3.2. Controllability rule-set

Based on the feedback from the Sweden workshop, separate tables for acceleration and deceleration were created. Additionally, to address the comment about over dimensionality, *"vehicle velocity"* was removed as a parameter. Table 8.8 and Table 8.9 depict the modified controllability rule-set.

Table 8.8: Controllability rule-set (part 1) for Germany workshop (workshop 3)

Emergency Deceleration Value	Distance to Obstacle	TTC	Controllability Rating
< 0.4g	< 6 m	< 1.0 sec	C3
		1.0 - 2.0 sec	C2
		> 2.0 sec	C1
	> 6 m	< 1.0 sec	C2
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0
Emergency Deceleration Value	**Distance to Obstacle**	**TTC**	**Controllability Rating**
0.4g - 0.8g	< 6 m	< 1.0 sec	C2
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0
	> 6 m	< 1.0 sec	C1
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0

Table 8.9: Controllability rule-set (part 2) for Germany workshop (workshop 3)

Acceleration Value	Distance to Obstacle	TTC	Controllability Rating
< 0.1g	< 6 m	< 1.0 sec	C2
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0
	> 6 m	< 1.0 sec	C1
		1.0 - 2.0 sec	C0
		> 2.0 sec	C0
Acceleration Value	**Distance to Obstacle**	**TTC**	**Controllability Rating**
0.1g - 0.4g	< 6 m	< 1.0 sec	C3
		1.0 - 2.0 sec	C2
		> 2.0 sec	C1
	> 6 m	< 1.0 sec	C2
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0

8.6.3.3. Exposure rule-set

Exposure rule-set has been illustrated in Table 8.10 (same as the one used in Sweden workshop (workshop 2).

Table 8.10: Exposure rule-set for Germany workshop (workshop 3)

Area	Driving Domain	Country	Exposure rating
City Centre	Pedestrian Pathways	India	E4
		Sweden	E1
		UK	E2
		Germany	E0
	Normal road	India	E4
		Sweden	E1
		UK	E2
		Germany	E1
Area	**Driving Domain**	**Country**	**Exposure rating**
Sub-urban areas	Pedestrian Pathways	India	E3
		Sweden	E0
		UK	E1
		Germany	E0
	Normal road	India	E4
		Sweden	E0
		UK	E1
		Germany	E0

8.6.4. Results

8.6.4.1. Quantitative results

Similar to workshop two, for every round, each of the groups were asked to provide Severity, Exposure and Controllability ratings for the two hazardous events (discussed in section 8.6.2.2). The ASIL ratings given by the four groups in different rounds have been illustrated in Figure 8.8. Four groups provided three different ASIL ratings in the first and the second round for hazard 1. Similarly, the four groups provided three different ASIL ratings for first round and two different ASIL ratings for second round for hazard 2. It is worthwhile to note that the difference in the ASIL ratings provided by the groups for hazard 1 was of the order of two (ASIL A – ASIL C) in both round one and round two. For hazard 2, the difference in ASIL ratings for round one was of the order of three (QM – ASIL C) and of the order of one (ASIL C – ASIL D) in round two. With the introduction of rules in round three, all groups converged to a QM ASIL rating for hazard 1. However, for hazard 2, a

variation of the order of one existed (QM – ASIL A) with two groups agreeing at QM rating and two groups giving an ASIL A rating. In round four, for hazard 1, the mixing of groups introduced a slight variation with one of the three groups giving an ASIL A rating while other three groups gave a QM rating. A similar variation was observed for hazard 2. It is interesting to note that it was the same group which gave the ASIL A rating to both the hazards in round 4.

Similar to workshops one and two, the large variation in the ASIL ratings given by the four groups in rounds one (order of 3 and order of 4 for hazard 1 and hazard 2 respectively) and round two (order of 3 and order of 1 for hazard 1 and hazard 2 respectively), emphasises the reliability challenges with automotive HARA process due to inter-rater variation (as discussed in chapter 3 and section 8.1.

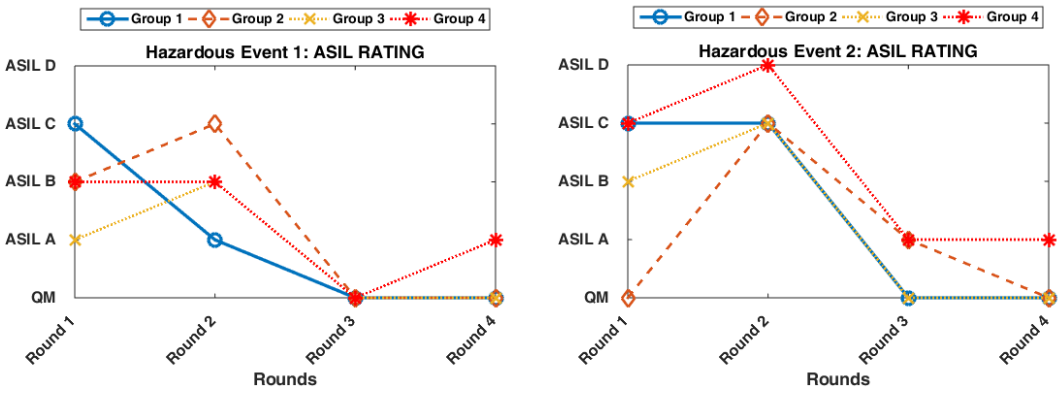

Figure 8.8: ASIL ratings for workshop 3 (Germany)

Severity

Figure 8.9 illustrate the severity ratings. In the absence of rules in round one, severity ratings showed a significant variation for hazard 2 (order of three (S0 and S3) in round 1) and for hazard 1 (order of one (S2 – S3). While in round two, severity ratings for hazard 1 had a

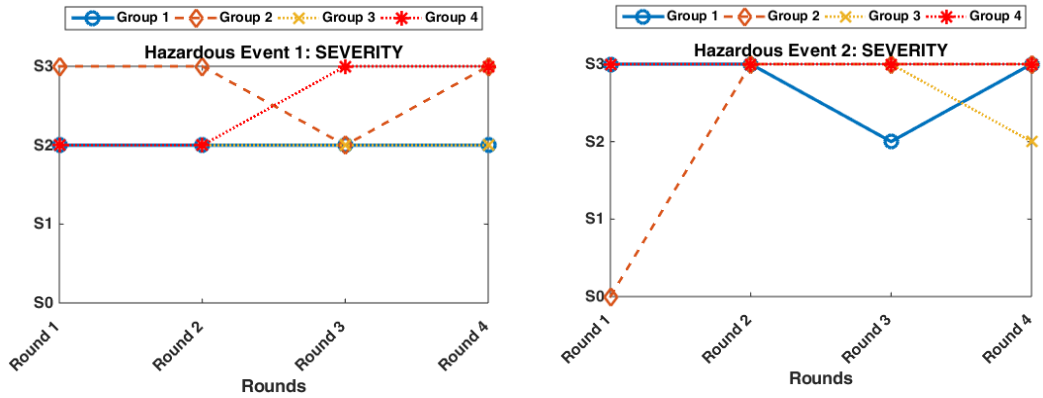

Figure 8.9: Severity ratings for workshop 3 (Germany)

variation of the order of one (S2 – S3), the ratings converged to S3 for hazard 2. With the introduction of rules in round three, while three groups converged in their ratings for hazard 1 (S2) and hazard 2 (S3), one group gave a different rating (S3 for hazard 1 and S2 for hazard 2). It is worthwhile to note that these were two different groups. The reasons for the variation even after the introduction of the rules are evident when the qualitative analysis of the results is considered which is discussed in section 8.6.4.2.

Exposure

In round one, exposure ratings for hazard one differed by the order of one between the groups (E3 – E4). Similar difference was observed in round two again for hazard 1. However, for hazard 2, a significant difference of the order of three was observed in round one (E1 – E4) with four groups giving three different ratings (E4, E3 and E1). In round two, a difference of the order of one (E3 – E4) was observed for hazard 2. With the introduction of rule-set, the exposure ratings for both round three and round four converged to E1 for hazard 1 and hazard 2 (Figure 8.10).

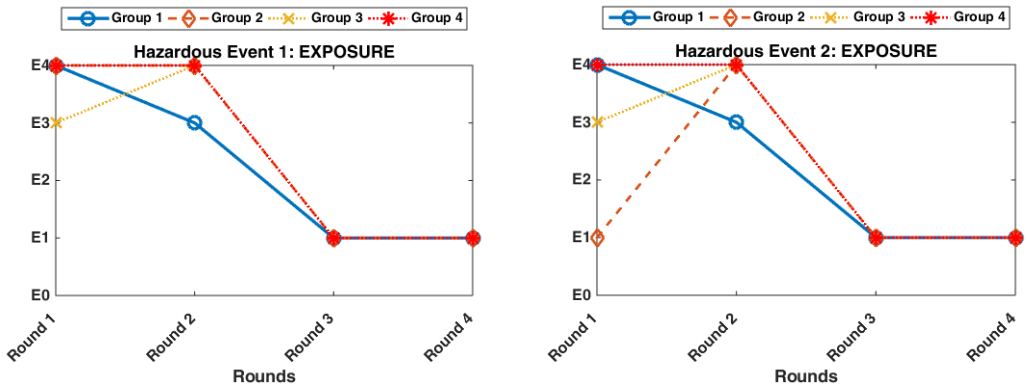

Figure 8.10: Exposure ratings for workshop 3 (Germany)

Controllability

As in the case with workshop 1 and workshop 2, controllability ratings observed the maximum variation, even with the introduction of the rule-set. In round one, controllability ratings provided by the groups differed by an order of two for both hazard 1 (C1 – C3) and hazard 2 (C0 – C2). For hazard 1, the three different ratings were observed (C1, C2 and C3). In round two, while the groups converged to C2 rating for hazard 1, a difference of the order of one (C2 – C3) was observed for hazard 2. The introduction of the rule-set in round three didn't improve the convergence in ratings as a difference of the order of one was observed in the controllability rating for both hazard 1 and hazard 2, which increased to the order of three in round four (Figure 8.11). The cause of variation even with the introduction of rule-

set in round three and four can be better explained with the qualitative analysis of the feedback received by the safety experts who took part in the workshop. The qualitative analysis of the feedback and the notes made by the moderators has been discussed in section 8.6.4.2.

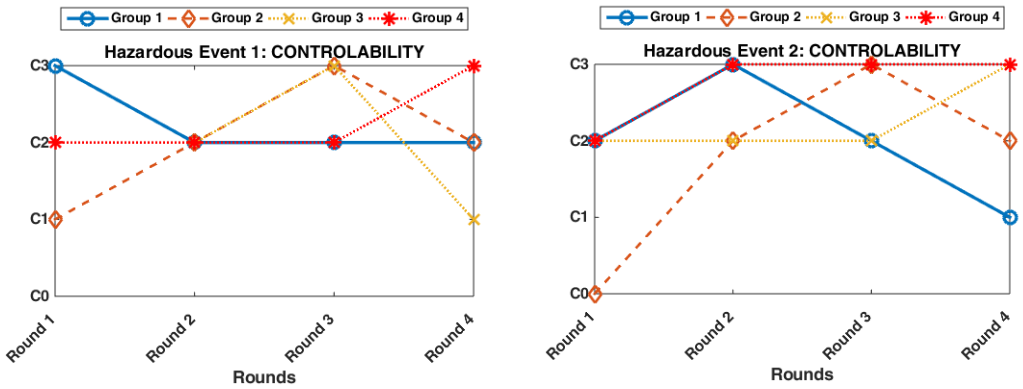

Figure 8.11: Controllability ratings for workshop 3 (Germany)

8.6.4.2. Qualitative results

In response to the question on whether the groups experienced variation in HARA in the various rounds, all groups agreed on the existence of variation in the ratings in the different rounds. Like in earlier workshops, most of the groups mentioned that the variation was caused due to difference in assumptions. One of the groups commented: "variation with all assumption…no stats (statistics)", while the other commented *"high variation due to non-normalised assumptions"*. Interestingly, none of the groups mentioned that experience in conducting HARA was one of the causes. This is because the experience level of above the participants was over 10 years.

When asked whether the introduction of rule-set by parametrising the HARA elements increased the reliability of the process, almost all the participants responded positively, with one of the participants suggesting: *"yes, puts some better context of the situation"*. Another participant mentioned: *"Yes, I think so, otherwise the procedure is really subjective"*. Interestingly, one of the participants commented: *"(parametrisation of HARA increases reliability) perhaps for severity, and controllability: if we know more about the vehicles. For HE2, we don't know about the vehicle, so had to go worst case. For exposure, I think this would help"*. While one of the participants had reservations about the concept suggesting: *"such an approach may lead to pseudo objectivity"*. This concern resonates with one of the comments from workshop 2, which suggested the possibility of adapting the HARA ratings

to fit the rule-set. This comment re-emphasizes the need to have an exhaustive rule-set in order to allay such fears.

8.6.5. Learnings from workshop 3: A discussion

While good convergence was reached for exposure ratings with the introduction of the exposure rule-set, participants mentioned the need to introduce an additional parameter for *"time of the day"* to make the rule-set more exhaustive. In workshop three, the reliability of the controllability ratings continued to be low (even with the introduction of the rule-set). One of the main reasons suggested by participants for the lack of convergence on the controllability ratings was the definition of the hazard and the hazardous events. Therefore, for the workshop four (final workshop), an even more detailed description of the hazard and hazardous event was used.

8.7. Workshop 4 (U.K.)

8.7.1. Participants

Eleven participants with extensive functional safety experience were recruited for the final workshop study. Each participant had over 5 years experience in safety analysis, with some having over 20 years experience. All participants were based in the UK. In addition to the 11 participants, three moderators were also recruited from the author's research group. Out of the 11 participants, four participants were from Original Equipment Manufacturers (OEMs), six were from Tier 1 suppliers and one participant was from an engineering consultancy. Except one, all participants identified themselves as either functional safety specialists or functional safety engineers.

8.7.2. Workshop structure

Eleven participants were grouped into three groups with two groups having four participants and one group having three participants. In addition to the eleven participants, each group was assigned a moderator. Like in earlier workshops (discussed in sections 8.4 (USA), 8.5 (Sweden) and 8.6 (Germany)) each group was provided with two different hazardous events and were asked rate the severity, exposure and controllability for the two given hazardous events.

Workshop four had a different structure to the earlier workshops. As workshops one, two and three had already established the existence of the inter-rater variation in automotive HARA, the author wanted to focus workshop 4 solely on the impact of the introduction of the rule-set in the automotive HARA process. Additionally, unlike other workshops where a

conference or a meeting gathering was used as a platform to gather various functional safety experts, workshop four involved inviting functional safety experts to the University of Warwick for the purpose of the workshop only. In order to make best use of the time provided by the experts, the author had only two rounds in workshop four. An extended feedback session was designed for workshop four, as it was the final workshop of the series. It was expected that the feedback session would provide additional insights into the challenges of the automotive HARA, the proposed process and the utility of the proposed process in an industrial context. The workshop structure for workshop four has been illustrated in Figure 8.12.

Figure 8.12: Structure for workshop 4

8.7.2.1. Additional System definition

In addition to the system definition for the SAE Level 4 Low-Speed Automated Automated Driving system (LSAD) described in section 8.3.3.1, participants were also provided with the following description:

- Maximum acceleration of the LSAD = 0.4g
- LSAD equiped vehicle dimensions: 2m x 2.5m x 2.5m
- Vehicle crashworthy upto top speed
- Driving domain: public roads in inner city in UK

The additional information was provided as it was received as a part of the feedback from workshop one, workshop two and workshop three.

8.7.2.2. Hazard and hazardous event definition

The hazard provided to the participants was *"unintended acceleration"*. Based on the hazard, participants were provided two hazardous events and were asked to discuss the HARA for the two given events to give Severity, Exposure and Controllability ratings. The two hazardous events provided to the participants were:

- Pod travelling around a corner at 15 kmph towards a pedestrian (child running) in an urban environment
- Pod crossing a junction at 15 kmph which is in the path of an oncoming vehicle

8.7.3. Rule-set

8.7.3.1. Severity rule-set

Severity rule-set has been depicted in Table 8.11, Table 8.12 and Table 8.13 (same as the one used in Sweden and Germany workshops (workshop 2 and 3).

Table 8.11: Severity rule-set (part 1) for UK workshop (workshop 4)

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Type of Impact	Severity Rating
Pedestrian (Adult)	< 11 km/h	< 2 km/h	Head-on	S0
			Rear	S0
			Side	S0
		< 6 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 12km/h	Head-on	S1
			Rear	S1
			Side	S1
	11 - 16 km/h	< 2 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 6 km/h	Head-on	S2
			Rear	S1
			Side	S1
		< 12km/h	Head-on	S2
			Rear	S1
			Side	S2
	> 16 km/h	< 2 km/h	Head-on	S2
			Rear	S1
			Side	S2
		< 6 km/h	Head-on	S3
			Rear	S2
			Side	S2
		< 12km/h	Head-on	S3
			Rear	S2
			Side	S3

Table 8.12: Severity rule-set (part 2) for UK workshop (workshop 4)

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Type of Impact	Severity Rating
Pedestrian (Child)	< 11 km/h	< 2 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 6 km/h	Head-on	S1
			Rear	S1
			Side	S1
		< 12km/h	Head-on	S1
			Rear	S1
			Side	S1
	11 - 16 km/h	< 2 km/h	Head-on	S2
			Rear	S1
			Side	S2
		< 6 km/h	Head-on	S2
			Rear	S2
			Side	S2
		< 12km/h	Head-on	S3
			Rear	S2
			Side	S3
	> 16 km/h	< 2 km/h	Head-on	S3
			Rear	S2
			Side	S2
		< 6 km/h	Head-on	S3
			Rear	S2
			Side	S3
		< 12km/h	Head-on	S3
			Rear	S3
			Side	S3
Type of Obstacle	**Vehicle Velocity**	**Oncoming Obj. Velocity**	**Type of Impact**	**Severity Rating**
Infra-structure	< 11 km/h	0 km/h	Head-on	S0
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
	11 - 16 km/h	0 km/h	Head-on	S1
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
	> 16 km/h	0 km/h	Head-on	S2
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	
		0 km/h	Head-on	
			Rear	
			Side	

Table 8.13: Severity rule-set (part 3) for UK workshop (workshop 4)

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Type of Impact	Severity Rating
Cyclist	< 11 km/h	< 8 km/h	Head-on	S0
			Rear	S0
			Side	S0
		< 14 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 20km/h	Head-on	S2
			Rear	S1
			Side	S2
	11 - 16 km/h	< 8 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 14 km/h	Head-on	S2
			Rear	S1
			Side	S2
		< 20km/h	Head-on	S2
			Rear	S2
			Side	S2
	> 16 km/h	< 8 km/h	Head-on	S2
			Rear	S1
			Side	S2
		< 14 km/h	Head-on	S2
			Rear	S2
			Side	S2
		< 20km/h	Head-on	S3
			Rear	S2
			Side	S3
Type of Obstacle	**Vehicle Velocity**	**Oncoming Obj. Velocity**	**Type of Impact**	**Severity Rating**
Vehicle	< 11 km/h	< 10 km/h	Head-on	S0
			Rear	S0
			Side	S0
		< 20 km/h	Head-on	S1
			Rear	S0
			Side	S1
		> 20km/h	Head-on	S2
			Rear	S1
			Side	S2
	11 - 16 km/h	< 10 km/h	Head-on	S1
			Rear	S0
			Side	S1
		< 20 km/h	Head-on	S1
			Rear	S1
			Side	S1
		> 20km/h	Head-on	S2
			Rear	S2
			Side	S2
	> 16 km/h	< 10 km/h	Head-on	S1
			Rear	S1
			Side	S1
		< 20 km/h	Head-on	S2
			Rear	S2
			Side	S2
		> 20km/h	Head-on	S3
			Rear	S2
			Side	S3

8.7.3.2. Controllability rule-set

Controllability rule-set has been depicted in Table 8.14 and Table 8.15 (same as the one used in Germany workshop (workshop three).

Table 8.14: Controllability rule-set (part 1) for UK workshop (workshop 4)

Emergency Deceleration Value	Distance to Obstacle	TTC	Controllability Rating
< 0.4g	< 6 m	< 1.0 sec	C3
		1.0 - 2.0 sec	C2
		> 2.0 sec	C1
	> 6 m	< 1.0 sec	C2
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0
Emergency Deceleration Value	**Distance to Obstacle**	**TTC**	**Controllability Rating**
0.4g - 0.8g	< 6 m	< 1.0 sec	C2
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0
	> 6 m	< 1.0 sec	C1
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0

Table 8.15: Controllability rule-set (part 2) for UK workshop (workshop 4)

Acceleration Value	Distance to Obstacle	TTC	Controllability Rating
< 0.1g	< 6 m	< 1.0 sec	C2
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0
	> 6 m	< 1.0 sec	C1
		1.0 - 2.0 sec	C0
		> 2.0 sec	C0
Acceleration Value	**Distance to Obstacle**	**TTC**	**Controllability Rating**
0.1g - 0.4g	< 6 m	< 1.0 sec	C3
		1.0 - 2.0 sec	C2
		> 2.0 sec	C1
	> 6 m	< 1.0 sec	C2
		1.0 - 2.0 sec	C1
		> 2.0 sec	C0

8.7.3.3. Exposure rule-set

Based on the feedback received during the Germany workshop, an additional parameter of *"time of day"* was added to the exposure ratings. Table 8.16 illustrates the modified exposure rule-set.

Table 8.16: Exposure rule-set for UK workshop (workshop 4)

Type of Obstacle	Area	Driving Domain	Country	Time of Day	Exposure rating
Pedestrians	City Centre	Pedestrian Pathways	India	Day	E4
				Night	E1
			Sweden	Day	E3
				Night	E1
			UK	Day	E3
				Night	E2
		Normal road	India	Day	E2
				Night	E1
			Sweden	Day	E3
				Night	E1
			UK	Day	E3
				Night	E1
	Sub-urban areas	Pedestrian Pathways	India	Day	E3
				Night	E2
			Sweden	Day	E2
				Night	E1
			UK	Day	E2
				Night	E1
		Normal road	India	Day	E4
				Night	E3
			Sweden	Day	E2
				Night	E1
			UK	Day	E2
				Night	E1
Type of Obstacle	**Area**	**Driving Domain**	**Country**	**Time of Day**	**Exposure rating**
Vehicle	City Centre	Pedestrian Pathways	India	Day	E3
				Night	E1
			Sweden	Day	E4
				Night	E1
			UK	Day	E4
				Night	E1
		Normal road	India	Day	E4
				Night	E4
			Sweden	Day	E4
				Night	E2
			UK	Day	E4
				Night	E2
	Sub-urban areas	Pedestrian Pathways	India	Day	E4
				Night	E2
			Sweden	Day	E3
				Night	E1
			UK	Day	E3
				Night	E1
		Normal road	India	Day	E4
				Night	E4
			Sweden	Day	E4
				Night	E2
			UK	Day	E4
				Night	E2

8.7.4. Results

8.7.4.1. Quantitative results

For the hazardous event 1, the three groups came up with three different ASIL ratings in round 1 (without rule-set) after the discussion, once again suggesting the existence of intra-rater variation. The ratings differed by an order of two (ASIL B – ASIL D). With the introduction of the rule-set in round two, the ASIL ratings for the three groups converged to ASIL C for hazardous event 1(Figure 8.13). While the convergence in ASIL ratings might indicate increased reliability due to the introduction of the rule-set, a deeper analysis of the severity, exposure and controllability ratings (discussed later in this section) is required before drawing such a conclusion.

For hazardous event 2, the groups vary in their ASIL rating by an order of two (ASIL C – ASIL D) in round one. However, the variation increased to the order of two (ASIL B – ASIL D) in round two (with the rule-set). It is interesting to note that the results for the two hazardous events differed from each other (convergence for hazardous event 1 and divergence for hazardous event 2) (Figure 8.13). The observed difference is better understood once the individual S, E and C components are discussed and with the qualitative analysis of the moderators' notes (discussed in section 8.7.4.2).

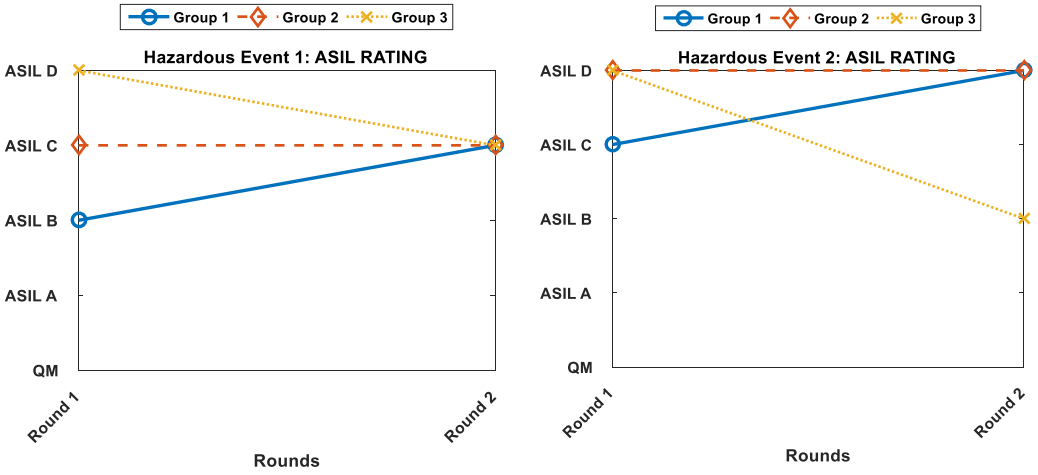

Figure 8.13: ASIL ratings for hazardous events 1 and 2 (workshop 4)

Severity

For round one, the severity rating differed by an order of one (S2 – S3) between the three groups for hazardous event 1. With the introduction of the rule-set in round two, the severity ratings showed a convergence to S3, indicating higher reliability due to the introduction of

the rule-set (Figure 8.14). However, the trend was different for hazardous event two. In both round one and round two (with the rule-set), the severity ratings showed a variation of the order of one for the hazardous event 2 (Figure 8.14).

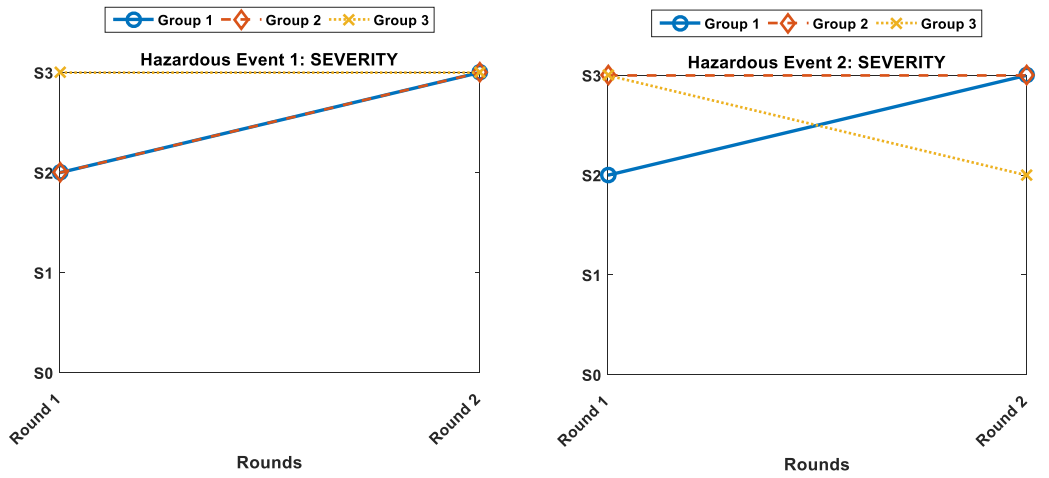

Figure 8.14: Severity ratings for hazardous event 1 and 2 (workshop 4)

Exposure

For hazardous event 1, exposure ratings showed a variation of the order of one (E3 – E4) in round one. With the introduction of the rule-set, the ratings converged to E3, indicating increased reliability with the introduction of the rule-set (Figure 8.15). For hazardous event 2, the exposure ratings were consistent at E4 between round one and round two (Figure 8.15).

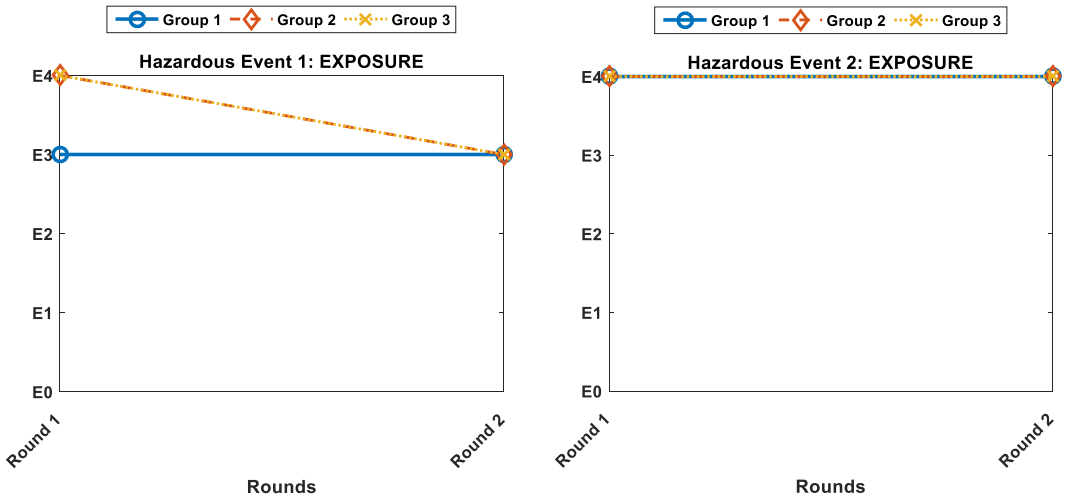

Figure 8.15: Exposure ratings for hazardous event 1 and 2 (workshop 4)

Controllability

Controllability ratings in workshop 4 showed a different trend as compared to the earlier workshops. Interestingly, UK experts seem to have been much more conservative in controllability ratings as compared with experts in earlier workshops. For hazardous event 1, the controllability ratings were consistent at C3 among all three groups for both round one and round two (Figure 8.16). However, for hazardous event 2, the controllability ratings for round one converged at C3, but showed a variation of the order of one (C3 – C2) in round two (Figure 8.16).

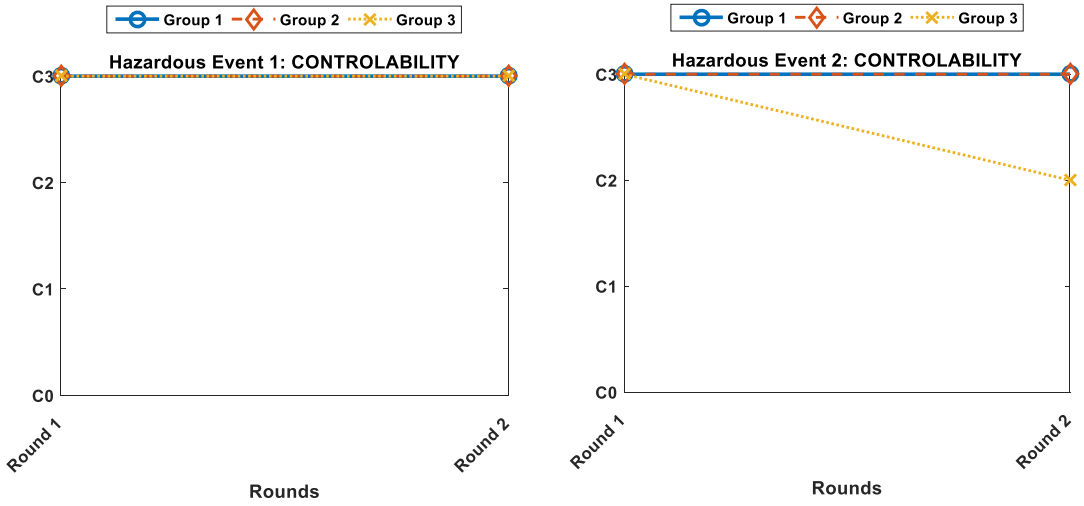

Figure 8.16: Controllability ratings for hazardous event 1 and 2 (workshop 4)

8.7.4.2. Qualitative results

Since in workshop four there were only two rounds and groups were not mixed, participants were not asked if they had experienced variation in the HARA ratings as it was not applicable. However, participants were asked to answer the second question (mentioned in section 8.3.3.3), about whether the introduction of rule-set by parametrising the HARA elements had increased the reliability of the process. The participants unanimously agreed that it would be possible to parametrise severity and controllability. However, they had their reservations about the ability to parametrise exposure ratings. One of the participants commented: *"this is possible and could be used. (But) rationale behind the table is very important"*. Another participant mentioned: *"(this process) could be used for new engineer for training"*. The participants suggested that the parametrised approach could be used to create a tool which aids the HARA process and the rationale could be incorporated within the tool.

Interestingly, it was observed that functional safety experts in the UK were conservative in their ratings, especially for controllability rating as it was always C3.

8.7.5. Learnings from workshop 4: A discussion

When asked about the challenges of the proposed approach (i.e. parametrisation of HARA elements to create a rule-set), the participants (like earlier workshops) mentioned that the rule-set needs to be exhaustive. The participants suggested adding more granularity to the rule-set by incorporating new parameters like *"shared space"* for controllability, *"population density"* for exposure and *"kinetic energy"* for severity. Some of the variation in ratings experienced even with the introduction of the rule-set could be attributed to these missing parameters and different groups making different assumptions, which can be seen in the variation in severity and controllability ratings (Figure 8.14 and Figure 8.16). In addition, they mentioned that one of the major challenges with HARA of ADAS or ADS is item definition.

8.8. Learnings from the workshop series: A discussion

One of the potential future benefits of having an objective rule-set is that it paves the way for dynamic HARA. With the introduction of automated systems, a concept of dynamic HARA has been introduced recently to enable the automated system to determine its ASIL rating based on the situational health of the sensors and the automated system and the environmental conditions (Johansson and Nilsson, 2016a; Villa et al., 2016). The approach presented in this chapter constitutes one of the blocks of a dynamic HARA and may aid in a reliable hazardous event rating in the dynamic HARA process (Johansson and Nilsson, 2016b). Additionally, it can potentially allow relatively unskilled practitioners with less experience, to perform HARA to a reliable degree as the need for highly specialized knowledge is reduced to a great extent. This could ease the process in terms of time and resources required for the HARA.

An objective approach to the decision making process involved in an automotive HARA has many potential benefits. Not only does it have the potential to increase the inter-rater reliability of the process, it provides the ability to automate the HARA process which in turn can save precious time in the automotive product life-cycle. Moreover, it can potentially provide a degree of consistency across the automotive supply chain (i.e., OEMs, Tier 1 suppliers, Tier 2 suppliers etc.). While some of the results suggest positive results towards increased reliability through convergence of HARA ratings, it is not known that full

convergence would ever occur but this work has shown that introduction of an objective rule-set has a potential to increase the reliability of HARA ratings.

When considered in the context of *"informed safety"* for the user of ADAS and ADS, the proposed objective HARA approach increases the reliability of the risk conveyed to the drivers which forms their *"informed safety"* level. It aids in the development of trust with the system as reliability of risk ratings increases predictability and increases drivers' ability to work with the system.

8.9. Conclusion

A novel approach to improve the reliability of the automotive HARA process has been proposed by creating a rule-set for each of the three HARA elements (severity, exposure and controllability). The proposed objective HARA approach has a potential to mitigate inter-rater variations caused by subjective nature of the functional safety experts' mental models and background knowledge.

The rule-set for HARA involves parametrization of the HARA parameters elements, i.e., severity, exposure and controllability. While discussing some of the benefits of the proposed objective HARA approach, the low reliability, i.e. intra-rater variation, of the current automotive HARA process has also been demonstrated in the four workshop studies discussed in this chapter.

Based on the feedback from participants and the qualitative analysis of results in each workshop, the rule-set was re-calibrated and the hazardous event definition was modified. One of the themes that was observed in the qualitative analysis of the feedback was the need to put a context to the hazard in the HARA. The perception of severity, exposure and controllability varies in different contexts. Additionally, it was also observed that there is a need to provide a user guidebook / user training on how to use the rule-set. The introduction of rule-set did suggest increased convergence, especially for severity and exposure ratings. Although, introduction of the rule-set also suggested increased reliability of controllability ratings, the effect was limited in comparison to severity and exposure.

The four workshops provided some positive results along with some mixed results. One of the reasons for the mixed results could be the need for more elaborate rule-set incorporating more parameters for severity, exposure and controllability. Future research may be carried out to extend the rule-set by incorporating more parameters and also experimentally validating the rule-set itself.

While full convergence in ratings between groups wasn't achieved at the end of four workshops, the participants in the workshop acknowledged that the proposed method could be used but a rationale for the rule-set needs to be provided to the experts. Additionally, in one of the workshops (workshop four), it was suggested that the objective HARA method would be useful to train new engineers. Participants in all the workshops felt that the introduction of the rule-set helped remove subjectivity of the HARA process and thus making the process more objective.

UNDERSTANDING RESULTS: A DISCUSSION

Chapter 9

In chapter 2, the author discussed that the adoption of ADASs and ADSs could offer many potential benefits like improved safety by reducing human error, lower emissions, increased traffic through-put, providing more productive time to the drivers, among many others. However, also introduced were some of challenges that need to be overcome in order to realise the benefits of ADASs and ADSs. In this thesis, two of the challenges highlighted in chapter 2 have been discussed in detail. These being: 1) reaping the safety benefits and 2) establishing the safety level.

Based on the two challenges, three specific research questions were identified based on the review of literature in chapter 3 and chapter 6. This chapter discusses the results from the research presented in this thesis and their wider contribution and highlights some of the solutions to the challenges mentioned above. This chapter is structured in three sections. Section 9.1 reflects on the research conducted to answer the three research questions, section 9.2 discusses the potential impact of the results in a wider research and industrial context, and finally, section 9.3 discusses the limitations of the research presented in this thesis and the potential for future work.

9.1. Reflection on the research

9.1.1. Reflection on the two themes of this thesis

This thesis has been structured under two main themes: 1) trust (chapter 3 and chapter 5) 2) testing (test scenarios and risk classification) (chapter 6, chapter 7 and chapter 8). While the two distinct themes exist in this thesis, they are brought together by the need for creating the *"content of knowledge"* in order to create a state of *"informed safety"*. This is required to calibrate trust to appropriately high levels. In order to calibrate trust to high levels, knowledge about the true capabilities and limitations needs to be provided to the driver. However, in order to create this knowledge, ADASs and ADSs need to be tested to establish their limitations and capabilities. This subtle but important link has been shown to be crucial in recent accidents involving the Tesla Model S (NHTSA, 2017a).

In one of the fatal accidents, the driver of the Tesla Model S assumed the "Autopilot" feature to be a fully-automated system and didn't monitor the system and the road, leading to the system colliding with a rear end of a truck. In the other accident, the Tesla Model S collided with a parked fire-truck with the driver once again expecting the "Autopilot" system to behave as a fully automated system. While in the former, the situation encountered by the Model S was one that wasn't tested (i.e., the limitation of the system wasn't established), the driver was either misinformed about the true capabilities and limitations of the system or had over-trust in the system's capabilities. In the case of the latter accident, the radar sensors on the Tesla were not able to detect the stationary fire truck (another case where the limitation of the system wasn't established). Both these incidents demonstrate the need to calibrate the trust level of the driver to prevent disuse or misuse of automation, to ensure safe use of ADASs and ADSs.

9.1.2. Reflection on the results

9.1.2.1. Reaping the safety benefits

Chapter two discusses that in order to realise the benefits offered by ADASs and ADSs, it is essential to ensure that drivers use such systems in their daily commutes. In chapter 3, the review of literature suggested that drivers' use of ADASs and ADSs is influenced by their trust in such systems. Trust is classified into: *"trust in the system"* which refers to the capability of the system and *"trust with the system"* which refers to the ability of the driver to work with the system.

Various peer-review comments were received on the concept "trust in the system" and "trust with the system" when the research from study 1 (trust and static knowledge) was submitted

for journal publication (Transportation Research Part C: Emerging Technologies). It was evident from the peer review comments that there was a need for a greater understanding of the concepts of two types of trust and the distinction between them. In order to address the review comments and to add clarity to the difference between the two forms of trust, the definitions were improved. This was also the case for other concepts like *"static knowledge"* and *"internal mental model"*. The peer review process achieved two main objectives. Firstly, it ensured that the new concepts like *"trust in the system"*, *"trust with the system"* and *"static knowledge"*, and their subtle differences was understandable to others. Secondly, it also highlighted the need for clear definitions for concepts.

In order to increase drivers' use of ADASs and ADSs, the author proposed to increase drivers' trust (both trust-in and trust-with) in ADASs and ADSs. In this thesis, it has been experimentally demonstrated that by providing knowledge to the drivers about the true (experimentally validated) capabilities and limitations of the system, their trust is calibrated to high levels, to increase the use of ADASs and ADSs, while also preventing disuse and misuse of the systems.

In chapter five, while discussing study 1 (where static knowledge about the true capabilities and limitations of the system was provided), it was concluded that *"trust in the system"* could be increased ($p < 0.001$) by introducing static knowledge. Interestingly, providing static knowledge about the true capabilities and limitations of the system didn't have any effect on *"trust with the system"* ratings ($p > 0.05$). This finding is especially interesting, considering some of the post-study run feedback received from the participants. One of the participants mentioned: *"warnings from the car missing"*, while another participant mentioned: *"no warnings & notification"*. The feedback suggested that the lack of real-time (dynamic) knowledge led to participants' inability to work with the system, thus causing no change in their *"trust with the system"* ratings.

Subsequently, study 6 in which dynamic knowledge about the true capabilities and limitations of the ADASs and ADSs was provided to the drivers (via HMI display), demonstrated that dynamic knowledge indeed has a larger positive effect on trust ratings (both *"trust in the system"* and *"trust with the system"*) ($p < 0.001$). Some of the qualitative feedback was contrary to the feedback received in study 1. One of the participants mentioned: *"I know when I have to be more alert"*, while another participant mentioned: *"visual information (red, amber or green) provides a measure of how alert driver (I) should be"*. The feedback suggests higher ability of the participants to work with the system with the introduction of dynamic knowledge.

For low capability automation, introduction of static knowledge led to a significant increase in workload experienced by the participants. However, introduction of dynamic knowledge negated this effect. Moreover, introduction of knowledge (static or dynamic) showed no significant change in workload for high-capability automation.

Effect of automation capability (low or high) in trust ratings in run 1 (no knowledge) and run 2 (with knowledge) provided another interesting insight. When static knowledge was introduced in study 1 (run 2), automation capability didn't have any significant effect on *"trust in the system"* ratings, suggesting similar trust ratings irrespective of the automation capability. Similar effect was noticed when dynamic knowledge was introduced in study 6 (run 2), automation capability had no interaction effect with dynamic knowledge for both *"trust in the system"* ratings and *"trust with the system"* ratings.

The two studies (in chapter 5), when considered together help establish that dynamic knowledge has a higher ability to positively affect trust (trust in the system and trust with the system) as compared to static knowledge.

Answering the Research Question (RQ) 1 (*"How to increase "trust in/with" automation in vehicles"*) (identified in chapter 3), has helped to provide an interesting avenue for increasing drivers' use of ADSs (preventing misuse and disuse) and thus potentially reaping the safety benefits offered by such systems. Furthermore, answering RQ 1 (in chapter 5), also suggests that it is possible to calibrate trust with the introduction of knowledge, even in a low capability automation system. Thus, suggesting that it is not essential to have a 100% full-proof ADASs or ADSs in order to have high or appropriate trust. However, it is important to create a state of *"informed safety"* where the drivers are aware of the true capabilities and limitations of the systems and hence prepared to work with the system to optimise performance. It is important to appreciate that the drivers' capability to overcome the limitations of the automated systems (even when they know about them) may differ based on the skill-set of the drivers. Their skill-set will decide the actions they take in a situation. While the research discussed in this thesis doesn't claim to have explored a representative sample of the population to capture relative differences in skills and attitude of people, this research discusses the relative difference in trust levels for an individual in the presence and absence of the state of *"informed safety"*. Section 9.3 on future work discusses the subject of differences between people and its implication on the results.

The state of *"informed safety"* will aid drivers' actions and which could potentially mean reduced use of automation (depending on situations) based on their own skill-set and their informed safety level. State of *"informed safety"* enables drivers to display knowledge-based behaviour by developing a high-level understanding of the working of the system. The

implication of this conclusion has potentially a larger impact on the speed of commercial introduction of ADSs (discussed in section 9.2.1).

It is worthwhile to note that the measure of trust is dependent on individuals and the difference between trust ratings observed in study 1 (trust and static knowledge) and study 6 (trust and dynamic knowledge) is a relative difference for each individual participant. While same research methodology was used for all experiment participants, the measure of trust was a subjective scale. While an objective measure of trust was not a part of the research in this thesis, the effect of knowledge (static and dynamic) on the relative difference in trust ratings and the concept of *"informed safety"* established by the work in this thesis, provides an answer to RQ 1: *how to increase trust "in/with" automation in vehicles*.

9.1.2.2. Establishing safety level

In chapter 5, it was established that by providing knowledge (static or dynamic) about the true capabilities and limitations of the system, drivers' trust could be calibrated to the appropriate levels by creating a state of *"informed safety"*. However, it is worthwhile to emphasize that the automated systems used in chapter 5 were dummy systems with known capabilities and limitations which were designed into the experimental setup.

For real-world systems, there is a need to create the knowledge of the true capabilities and limitations of ADASs and ADSs. This knowledge can be created by evaluating such systems in various test scenarios to establish their capability to tackle them. The author explored the process of how to create this knowledge (to be imparted to the driver) in chapter 6 by creating a method to identify test scenarios for testing.

Introduction of ADASs and ADSs in cars is turning them into complex systems with millions of lines of code. It is said that today's cars (even with limited automation) have over 100 million lines of code as compared to 6.5 million lines of code in a Boeing 787 airplane (Charette, 2009; Strandberg et al., 2018). As the complexity increases, the testing (verification and validation) processes to ensure the product is safe (both from a functionality perspective and consumer trust perspective), become more challenging, both from an engineering perspective and also consumer confidence point of view. To prove that ADASs and ADSs are safer than humans, it has been suggested that the vehicles need to be driven for a cumulative of 11 billion miles (Kalra and Paddock, 2016b). While it is evident that this number is infeasible, it is suggested that instead of focussing on the number of miles, testing needs to focus on scenarios experienced by the ADASs and ADSs in those miles (Wachenfeld and Winner, 2017a). Thus, research question 2 was identified as: *"How to create test scenarios to establish the limitations of the automated driving systems?"*

In chapter 7, analysis of a series of semi-structured interviews with automotive verification and validation experts from different countries (Germany, Sweden, USA, UK and India) conducted as a part of this thesis, concluded that for ADASs and ADSs, one needs to test *"how a system fails"* rather than *"how a system works"*. Thus, suggesting that confidence in testing should be obtained not by the number of miles driven but the quality of miles driven. Based on these findings, a two-branched approach to testing was proposed, 1) Requirement Based Testing (RBT) and 2) Hazard Based Testing (HBT). While requirement based testing has been an integral part of the automotive development process, hazard based testing approach is a new addition to the testing methodologies within the automotive domain. The proposed hazard based testing comprises of three aspects, 1) hazard identification, 2) creating test scenarios from the identified hazards and 3) risk analysis for the identified hazards.

In chapter 7, for hazard identification, Systems Theoretic Process Analysis (STPA) method has been chosen as it identifies more hazards as compared to other hazard identification methods. Inspired by a systems' engineering approach, STPA captures sub-system interactions and identifies the hazards associated with it. For creating test scenarios from hazards, the author proposed an extension to STPA. In order to create test scenarios: two levels of parametrisation are performed. Firstly, base parametrisation is conducted to create the scenery and the dynamic elements of a test scenario. Secondly, parametrisation of the elements of STPA output is done to create hazard based scenarios. In order to evaluate the applicability of the proposed method, it was then applied to a real-world system (low-speed automated driving system) and the results were discussed and presented in chapter 7.

Answering RQ 2 (*"How to create test scenarios to establish the limitations of the automated driving systems?"*) mentioned in chapter 1, provides a potential method for establishing the true capabilities and (more importantly) limitations of the ADASs and ADSs. As shown in chapter 5, failures during operation of ADSs have a detrimental effect on drivers' trust in such systems. Therefore, being able to identify *"how the systems fail"* not only helps establish the true performance capabilities of ADASs and ADSs, it also helps to create a state of *"informed safety"* for the drivers leading to high degree of appropriate trust. At the same time, *"informed safety"* enables to be better prepared to deal with the limitations of the ADASs and ADSs. While traditional RBT method covers only a fraction of the possible test scenario space for the systems, the second testing branch (HBT) improves the coverage of the test scenario space which can be considered as *"negative scenarios"* (Alexander, 2003). Thus HBT, by providing a structured method to test "how a system fails", is able to meet two of the qualities suggested by the automotive verification and validation experts in their

interviews: 1) ability to test safety goals and ways in which the system may fail or reach system limits 2) structured method of identifying the system limits or failure scenarios.

In chapter 8, the author discusses the reliability of the risk classification associated with the identified scenarios. Through a series of workshops involving functional safety experts from different countries, it was demonstrated that the current automotive Hazard Analysis and Risk Assessment (HARA) process suffers from inter-rater variation, lowering the reliability of the ratings. It is important to have reliable risk classification as it is part of the dynamic knowledge imparted to the drivers to calibrate their trust levels. One method to ensure that more reliable risk classification (i.e. convergence in risk ratings) is given for the same scenarios by different experts is to provide a set of rules for the HARA components (severity, exposure and controllability) leading to the objectification of the HARA ratings (proposed in chapter 8). However, full convergence was not achieved. Interestingly, while some of the qualitative feedback from the experts mentioned that full convergence may never be achieved, they suggested that an objective approach to the HARA process via the parametrisation of the HARA components is a good starting point for engineers who are new to the HARA process, or when classification for relatively less complex hazardous scenarios (in terms of number of actors and ODD) is to be done, or indeed can be used as a check list for things that should be considered. While this meant that RQ 3 (*"How to improve the inter- and intra-rater-reliability of the automotive HARA process"*) identified in chapter 6, wasn't fully answered, the results of the workshop studies in chapter 8 did suggest that the objective approach (with a more elaborate rule-set) could potentially improve inter-rater reliability of the automotive HARA process. Thus, opening avenues for future research.

The key contributions of this research in terms of knowledge are:

- Trust in ADASs and ADSs can be increased by creating a state of *"informed safety"*. The state of informed safety is created by conveying the true capabilities and limitations of the system to the driver. This can be done by providing *"static knowledge"* (e.g. through feature training at the time) or *"dynamic knowledge"* (e.g. through a HMI interface). The state of informed safety also prevents misuse and disuse of the automated systems.
- By creating a state of *"informed safety"*, it is possible to introduce imperfect automation as long as the imperfections are conveyed to the drivers. This will enable early realisation of the benefits of automated technologies, e.g. increased safety, lowered emissions etc.
- Hazard Based Testing (HBT) can be used to establish the knowledge of the true capabilities and limitations of ADASs and ADSs. HBT focuses on how a system

fails rather than how a system works. HBT can be performed by using Systems Theoretic Process Analysis (STPA) along with an extension to create test scenarios.

- Automotive Hazard Analysis and Risk Assessment (HARA) suffer from reliability issues. The reliability of automotive HARA can be improved by objectification of the severity, exposure and controllability ratings.

9.2. Potential Impact of the results

9.2.1. Reaping the safety benefits

One of the on-going debates in the research community and the industry about safety of ADASs and ADSs is *"how safe is safe enough"* to reap the benefits of the ADASs and ADSs. Kalra and Groves (2017) via a Model for Automated Vehicle Safety (MAVS) argue that a 10% safety performance improvement (in terms of number of fatalities per 100 million vehicle miles travelled) over human drivers can have a much larger benefit (both in short term and long term) in saving lives as compared to more stringent technical requirements of 75% safety performance improvement or 90% safety performance improvement over human drivers, due to the delayed introduction of the latter two levels of safety performance. This is inspired by the idea that it will take a much longer time to prove that the performance of ADASs and ADSs is 75% or 90% better than human drivers.

The results of study 1 (static knowledge and trust) and study 6 (dynamic knowledge and trust) (in chapter five) lend more support to the introduction of limited capability ADSs, provided that the knowledge about the true capability and limitations of the systems are conveyed effectively to the driver. A prompt introduction of ADASs and ADSs whose true capabilities and limitations having been established and conveyed to the drivers, could help realise some of the benefits that ADASs and ADSs have to offer at a much earlier stage.

In modern ADASs and ADSs the capabilities and limitations might change during the lifecycle of use due to Over The Air (OTA) upgrades. In such instances, it is essential to inform the driver of the new capabilities and limitations. It can also be suggested that the drivers might need to provide a confirmation that they have understood the new capabilities and limitations. The subject of how to convey the knowledge of the capabilities and limitations to the driver remains out of scope of this thesis. Nevertheless, the results from studies in chapter five also suggested that with the introduction of knowledge (static) about the capabilities and limitations of the automated system, the workload increases for a low-capability automated system. However, such an effect wasn't noticed for high capability automation or when dynamic (real-time) knowledge was provided to the participants. Thus,

lending support to the idea of introducing high capability automated system with knowledge about the capabilities and limitations being conveyed effectively to the driver via static and dynamic knowledge. Section 9.3.2 discusses some of the ways the research question about how to impart knowledge could be answered.

The concept of *"informed safety"* discussed in chapter five provides an interesting insight into the responsibility of manufacturers and engineers developing ADASs and ADSs. To convey true capabilities of the systems, it is essential that manufacturers and engineers provide an honest insight into the state of the current technology and honest predictions for future deployment to avoid creating media hype associated with automated driving technology. At the same time, by creating a state of *"informed safety"*, we can also achieve early adoption of ADASs and ADSs, allowing the society to reap benefits of the technology.

9.2.2. Establishing safety level

Another challenge in understanding *"how safe is safe enough"*, lies in identifying scenarios which illustrate the limitations of the ADASs and ADSs. It is interesting to note that the answer to this challenge has evaded both the research community (as shown in chapter 6) and the industry (as shown in chapter 7 via the semi-structured interview study). In answering RQ2, the author has created a method which can identify the limitations of a system and thus highlight its true capabilities. The proposed approach of hazard based testing made use of the STPA method for hazard identification (which has been proven to be superior to other hazard identification methods in the literature). The evaluation of the proposed method wasn't the explicit aim of this thesis. However, the research work in this thesis contributes towards the research gap of understanding how to create test scenarios for ADASs and ADSs efficiently (i.e. identifying the unknown knowns and the black swan scenarios), and overcoming the 11-billion mile challenge (Kalra and Paddock, 2016a). As discussed in section 9.2.1, knowledge of the limitations can potentially help reap the benefits of the ADASs and ADSs with their quicker commercial introduction.

The key contribution of this research is to *move the focus away from the number of miles required to drive ADASs or ADSs, to focussing on the quality of miles – miles which uncover failures in the systems. This is achieved with the concept of Hazard Based Testing (HBT).* The implementation of HBT in this thesis uses STPA as the hazard identification method. While there are other hazard identification methods such as FTA, FMEA, HAZOP etc., the results of STPA haven't been compared with these methods for the case study as a part of this thesis. Future work would potentially involve comparing the results of the various hazard identification methods.

9.2.2.1. Influencing ISO standards

While the research conducted in this thesis has been published in peer-reviewed journals and conference proceedings, some of the results are also influencing international standards. The research conducted in this thesis (on test scenario development) is influencing the content of a new international (ISO) standard on Low-Speed Automated Driving (LSAD) system (ISO NP 22737). In chapter 7, a novel method for test scenario identification was proposed. The case study for the proposal was an LSAD system. Some of the results from the case study, i.e., test scenarios identified from the proposed STPA inspired process have been incorporated in the LSAD ISO standard draft. In addition, the STPA process also led to identification of safety requirements which have informed discussions of the LSAD draft standard. The author is the work item leader for the proposed standard on LSAD and represents the UK on the ISO Technical Committee 204 Working Group (WG) 14. The LSAD work item currently includes functional requirements, performance requirements and test procedures for LSAD systems.

9.3. Future work

9.3.1. Limitations of the research presented

In this section, before discussing the potential avenues for future work, the author discusses some of the limitations of the research presented in this thesis which potentially could influence future work.

In the studies discussed in chapter 5 (study 1 and study 6: effect of knowledge on trust) and chapter 7 (study 3: controllability study), a driving simulator was used as an experimental setup. Driving simulators offer many advantages but at the same time have some disadvantages too. One of the benefits of a driving simulator is that it allows participants to be put in an immersive environment while letting them experience hazardous situations without any physical harm (Winter et al., 2012). Also, since there are no known SAE Level 4/5 and few SAE level 3 automated driving systems in market, a driving simulator environment enables us to create a setup that potentially resembles future systems. Literature suggests that driver behaviour in a manually driven car in a driving simulator environment is predictive for real-world environment. However, transferability of results on driver behaviour in automated driving systems in driving simulator to real-world is unclear. Driver behaviour using automated driving systems will potentially be influenced by their experience with such systems and may also be affected by the interaction with other road users. Therefore, transferability of the results to real world needs to be evaluated separately.

Driving simulator fidelity plays a key role in stimulating drivers' behaviour that is representative or predictive of their real world behaviour. Low fidelity simulators can potentially invoke unrealistic driving behaviour and thus invalidate the study results (Winter et al., 2012). While it is suggested that safety of the participants in a driving simulator experiment is one of the major benefits of driving simulators, it may also potentially prove to be a disadvantage. Since participants are not exposed to real danger or do not face the real consequences of their actions (e.g. accidents, near misses, vehicle damage etc.), there is a possibility that it gives them a false sense of safety or competence. In an attempt to add a sense of risk and consequence to the experiments conducted in chapter 5 and chapter 8, scoring criteria was used with participants being given bonuses and penalties on "good" and "bad" behaviour.

The three driving simulator studies (in this research) were conducted at WMG, University of Warwick. A higher proportion of the study participants were from the university who were either more inclined to accept new technology or familiar with new technology. While the participants represented a wide age group (23 – 65 years), being close to research and technology development at the university, would mean they are more aware of the technology. Future work could potentially involve repeating the studies (similar to study 1, study 2 and study 6) with participant demography who would be least likely to accept new technology. This could provide an interesting insight in the transferability and generalisation of the results presented in this thesis.

In this research, WMG's 3xD Simulator for Intelligent Vehicles was used. While, the 3xD simulator provides a 360° Field of View immersive environment, it provides limited motion feedback. Due to lack of representative motion feedback, there is a potential for the onset of simulator sickness symptoms amongst the participants. This effect was especially noticed in study 1 (trust and static knowledge) where the scenario design included roundabout and steep turns. The lack of realistic motion feedback around roundabouts and steep turns led to the onset of simulator sickness for some of the participants (leading to 9 out of 57 (15.78%) participant dropouts in study 1 and 3). While results from the participants who felt simulator sickness were removed from the analysis, this learning influenced the simulator scenario design in for study 6, which included only sweeping bends. Thus, minimizing the effect of lack of motion feedback in the 3xD simulator. Incorporating the learning in study 6 resulted in a reduction in the participant dropout rate due to simulator sickness from 15.78% to 5.4% (2 out 37 participants in study 6). Moreover, literature suggests that short sessions with rest breaks help evade simulator sickness (Winter et al., 2012). All driving simulator experiments conducted in this research followed this protocol to prevent the onset of simulator sickness.

In chapter 5, it was concluded that providing dynamic knowledge about the true capabilities and limitations of the system can lead to a positive effect on calibration of trust to appropriate levels. However, the manner of imparting knowledge (i.e., audio, audio-visual, haptic etc.) has not been discussed in this thesis. While existing research (Biondi et al., 2017; Dambock et al., 2013; Petermeijer et al., 2017) suggests multi-modal feedback is the most effective in order to increase drivers' awareness, its effect on trust, or use in automated vehicles, hasn't been evaluated.

In chapter 8, the author discussed the objectification of the automotive HARA in order to improve the reliability of the HARA process. The author created rule-sets for severity, exposure and controllability ratings by parametrising each of the aspects. While the results discussed in chapter 8 show increased convergence (with respect to inter-rater variations) in ratings (between different groups of experts), full convergence was not reached. The size of the parametrised rule-set to cover complex scenarios was limited. This was also noticed in the feedback from the experts who suggested addition of more parameters to the rule-set to further improve reliability of the process.

9.3.2. Potential next research steps

In chapter 5, the author evaluated the effect of knowledge on the development of trust. Prospective next research steps could include the evaluation of some of the other factors identified in chapter 3 (literature review on trust) on the development of trust and development of an objective measure of trust.

In chapter 7, the proposed method for test scenario creation using hazard based testing, was applied to an example real-world system and the scenario creation process was discussed via an example sub-system of the LSAD. Prospective next research steps could include for the proposed method to be applied to other types of ADASs and ADSs to further evaluate its effectiveness.

The following are the potential next research questions (as a result of the work presented in this thesis) that could be evaluated:

<u>In the trust theme</u>:

- How to objectively measure trust of an individual in/with automation?
- How to impart knowledge about the automation capability to the drivers to calibrate trust to appropriate high levels?

- How much knowledge (and in which form) about automation capability is the optimum level to be provided to the drivers to calibrate trust to appropriate high levels? How is this affected by the drivers' capabilities/skills?

Some research areas to objectively measure trust have been linked with biometric correlation. The link (or correlation if any) between biometric indicators like galvanic skin response, pupil dilation, heartrate etc. and subjective trust ratings could help establish a more objective way of measuring trust. One of the major challenges in using biometric indicators is around the reliability of the measurement devices. Another avenue of potential research for objective measure of trust is to use a hybrid metric combining different biometric indicators.

Knowledge to create an informed safety state, can be imparted to drivers/users of ADASs and ADSs in various ways. As discussed in chapter six and in section 9.2.1, static knowledge and dynamic knowledge can have potentially different effects on trust and workload. As suggested by study six (dynamic knowledge and trust), dynamic knowledge can have a substantial effect in increasing trust and reducing workload. Dynamic knowledge can be imparted using different modalities (e.g. visual cues, audio cues, haptic cues or multi-modal cues). Furthermore, each of the modalities can have different levels of aggressiveness. This area of research will require carefully designed driving simulator or real world experiments tacking one modality at a time with a good control for the experiment to evaluate the effects.

In the testing theme:

- Which parameters should be added (to proposed rule-set) to achieve full convergence of the HARA ratings?
- How to quantify the benefit of Hazard Based Testing over other existing testing methodologies?

To identify the parameters for the rule-set for severity, controllability and exposure ratings, the identified hazards need to be parametrised. The parameters influencing the hazards will then need to be incorporated in the rule-set. This suggests that the rule-set for the HARA will be hazard dependent. In order to create a robust rule-set, future workshop studies (similar to study 3 and 4) will need to be conducted using different hazards. Additionally, before the workshops, the hazards will need to be parametrised. This could potentially be done using the STPA method and the extension suggested in chapter 7.

In order to quantify the benefit of Hazard Based Testing, a case study needs to be conducted to apply STPA inspired HBT and traditional Requirements Based Testing. A comparison of the difference in the number and the types of scenarios will provide an insight in to the

benefit of HBT over traditional methods. Additionally, another area of future research would be to establish the most effective hazard identification method (among STPA, FTA, FMEA and HAZOP) for HBT through a real-world case study.

9.4. Summary

The research discussed in this thesis provides an insight in the process of development of trust in ADASs and ADSs. The results suggest that by providing knowledge (in real-time) about the true capabilities and limitations of the ADASs and ADSs, drivers' trust can be calibrated to high levels. This also prevents drivers' misuse or disuse of these systems, and thus enabling the society at large to reap the benefits of ADASs and ADSs in a safe manner. Additionally, this thesis also presents a novel method to create the knowledge (which can be used to calibrate trust) about the true capabilities and limitations of the systems by testing *"how the systems (ADAS and ADS) fail"* rather than *"how the systems work"*. Some of the research contributions of this thesis have influenced the development of future international standards, demonstrating the impact of the presented work.

CONCLUSIONS OF THE RESEARCH

Chapter 10

In this chapter, the findings and conclusions from the research work discussed in this thesis are summarised. Review of literature (in chapter 3 and chapter 6) led to the creation of the three research questions explored and answered in part in this thesis. While Advanced Driver Assistance Systems (ADASs) and Automated Driving Systems (ADSs) offer many benefits, their benefits can only be realised if drivers use such systems. In order to increase drivers' use of such systems, the following conclusions were reached from the review of the literature:

- Trust is a key factor influencing drivers' use of automated features like ADAS and ADS.
- It is essential to calibrate trust to the appropriate level to avoid misuse (due to over-trust) or disuse (due to under-trust)
- Trust can be quantitatively classified as *"trust in the system"* and *"trust with the system"*. This is an important distinction as both the constructs represent different aspects of trust development and are influenced differently by various factors influencing development of trust.

In order to answer research question 1 (**How to calibrate *"trust in/with"* automation in vehicles?)**, firstly it was established that trust is influenced by many factors like knowledge, certification, workload, situation awareness, self-confidence, experience and consequence. Subsequently, Research Objective 1 (*"To develop a conceptual model for development of*

drivers' trust in automated driving systems"), was met by creating a conceptual model for development of drivers' trust in automated driving systems (in chapter 3).

In the conceptual model created, trust could potentially be increased by a variety of factors using many intervention methods. In this thesis, knowledge as a factor was the focus of the research. In chapter 3, knowledge has been classified into three aspects: 1) static knowledge 2) dynamic (real-time) knowledge 3) internal mental model. The two former types of knowledge (static and dynamic) have been discussed in this thesis as it is possible to impart them as a part of system design.

In chapter 5, it was demonstrated "knowledge" can be used as an intervention method to calibrate (increase or decrease) drivers' trust. Research objective two (*"To evaluate the effect of knowledge (static and dynamic) on calibration of trust"*), was met via two driving simulator studies that were conducted to evaluate the effect of knowledge (static and dynamic) on trust.

In chapter 5, results from two driving simulator studies led to many conclusions. Firstly, *"Trust in the system"* ratings can be increased (statistically significantly) by providing static knowledge about the automated system's capabilities and the limitations. Secondly, with introduction of static knowledge, the difference in *"trust in the system"* ratings between low capability and high capability automation was statistically insignificant, suggesting that introduction of static knowledge leads to similar trust levels for both automation capabilities. However, the introduction of static knowledge caused a significant increase in the workload for low capability automation. For high capability automation, static knowledge didn't have any such effect. Thirdly, the introduction of static knowledge didn't have any significant effect on *"trust with the system"*. Fourthly, dynamic (real-time) knowledge about the system capabilities and limitations, significantly increases both *"trust in the system"* and *"trust with the system"* ratings for both high capability and low capability automation. Interestingly, providing dynamic knowledge had an insignificant effect on workload ratings for both types of automation. Moreover, contrary to static knowledge, with the introduction of dynamic knowledge, difference in trust ratings (trust in the system and trust with the system) between low capability and high capability automation was statistically significant. Therefore, it can be concluded that dynamic knowledge is more beneficial and has a larger effect on trust ratings.

Thus, research question 1 was answered by establishing that introduction of knowledge (static or dynamic) helps calibrate *"trust in the system"* and *"trust with the system"* to appropriate levels by creating a state of *"informed safety"*. Moreover, providing dynamic knowledge had a larger influence on trust as compared to static knowledge. Informed safety

means informing the driver (via static and dynamic knowledge) about the true capabilities and limitations of the ADASs or Automated Driving Systems (ADSs) and the real-time state of the ADASs and/or ADSs. By creating a state of "informed safety" there is a potential to achieve early adoption of ADASs and ADSs enabling the society to reap the benefits of the technology rather than waiting for perfect automation.

Once it was established that providing knowledge to drivers about the true capabilities of the ADASs and ADSs helps to calibrate trust to appropriate levels, the process of creation of knowledge was subsequently explored. Literature suggested that testing is a potential method of establishing the true capabilities and limitations of the ADASs and ADSs, which form the content of the knowledge to be imparted to the drivers. Testing comprises of test methods, test scenarios and safety analysis. In chapter 6, the challenges associated with the creation of test scenarios and safety analysis were identified from the literature review. The following conclusions were made from the literature regarding test scenarios and safety analysis:

- There is lack of a structured approach in the process of creation of test scenarios.
- There is a lack of understanding in defining the content and objective of test scenarios to identify the true capabilities and limitations of ADAS and ADS.
- Classification of risk associated with a scenario suffers from reliability challenges (both inter-rater and intra-rater reliability).

This lead to the creation of two research questions (RQ2 and RQ3):

- ***How to create test scenarios to establish the limitations of automated driving systems? (RQ2)***
- ***How to improve the inter- and intra-rater-reliability of the automotive HARA process? (RQ3)***

In chapter seven, research question two (***How to create test scenarios to establish the limitations of automated driving systems?***) was answered by meeting the two associated research objectives. Firstly, in order to meet research objective three (*"To develop an understanding about the characteristics of a test scenario (especially for ADAS and ADS)"*), a semi-structured interview study of experts was conducted. It was concluded that for ADASs and ADSs, it is necessary to test *"how the system fails"* in addition to *"how the system works"*. Thus, research objective three was met and *"hazard based testing"* (HBT) as a new approach to testing was created as a proposal for creating test scenarios.

Once hazard based testing was proposed to test *"how a system fails"*, the research objective four (*"To develop a methodology for creating test scenarios (based on the identified*

characteristics)"), was met by creating a methodology for generating test scenarios. To create test scenarios via HBT, firstly hazards need to be identified and secondly test scenarios need to be created for the identified hazards. In order to identify hazards, Systems Theoretic Process Analysis (STPA) was used as it can analyse the sub-system interactions and capture hazards associated with them in a more efficient manner as compared to other hazard identification processes. In order to generate test scenarios, elements of the STPA output were parametrised. The proposed process was applied to a real-world case study of a low-speed automated driving system (pod) and test scenarios based on the proposed approach were created.

Thus, research question two was answered by establishing the characteristics of a test scenario to establish the limitations of ADAS and ADS and creating a methodology for creating test scenarios which possess the identified characteristics.

In chapter eight, research question three (***How to improve the inter- and intra-rater-reliability of the automotive HARA process?***) was answered. In order to address the reliability challenges associated with automotive Hazard Analysis and Risk Assessment (HARA), the author created an objective approach to HARA. A set of rules for severity, exposure and controllability ratings was created which guide the HARA process. The rule-set introduces parametrisation of the severity, exposure and controllability ratings. Thus, research objective five (*"To develop a rule-set for conducting automotive HARA"*) was met.

In a series of workshops involving international functional safety experts, it was demonstrated that the introduction of the rule-set led to improved convergence in HARA ratings. However, further work is required on the rule-set by introducing more parameters to further increase convergence in the HARA ratings. Therefore, research objective six (*"To determine the ability of the developed rule-set for HARA in improving the reliability of the automotive HARA"*), was partially met as the rule-set needs further re-calibration.

Thus, research question three was partially answered by creating an objective approach to automotive HARA by parametrising the severity, exposure and controllability ratings to create a rule-set for each of them which led to a partial convergence in the HARA ratings. However, further work is needed on the rule-set as more parameters need to be included in the rule-set for full convergence. Subsequently, the effect of the rule-set needs to be evaluated again.

In conclusion, this thesis presents a number of contributions to the research community on increasing trust in ADASs and ADSs and how to test ADASs and ADSs. Some of the larger contributions to research and industrial community are as follows:

- It is not essential to have perfect automation which is 100% capable in order to have high trust. However, to develop high trust, it is essential that the driver has the knowledge about the true capabilities and limitations of systems.
- Moreover, the knowledge of true capabilities and limitations needs to be imparted to the driver in a manner that enables them to display Knowledge Based (KB) behaviour. In KB behaviour, the drivers have an understanding of the high-level working of the system, and understand the reasons for the limitations and their role in how to overcome the limitations.
- Many studies have suggested that there is a need to drive over 11 billion miles to establish the true capabilities and limitations (safety) of the ADASs and ADSs. In a departure from traditional way of testing where *"how a system works"* is tested, for ADAS and ADS it is more important to test *"how a system fails"*. Hazard Based Testing (HBT) has the potential of identifying scenarios where the system would fail.
- The knowledge about system limitations created through HBT can be imparted to drivers to create a state of *"informed safety"* allowing them to adapt their usage of the system to work with the system by overcoming its limitations.
- Systems Theoretic Process Analysis (STPA) inspired hazard based testing identifies hazards especially in complex systems like ADASs and ADSs which involve human-automation interaction. The proposed extension to STPA (parametrisation of STPA outputs) provides a structured approach to define test scenarios to establish the limitations of ADASs and ADSs.
- Having identified various scenarios, the next step in safety analysis is the classification of risk associated with the scenarios. The objective approach to hazard analysis and risk assessment (proposed in this research) has a potential to decrease the inter-rater and intra-rater variation in the risk ratings and also help train new engineers in the hazard analysis process by aiding their thought process on the parameters to be considered for risk analysis.
- From an industrial impact perspective, the STPA inspired hazard based testing has received interest from various industrial partners. Test scenarios resulting from the

proposed process are being considered for incorporation in international standard on Low-Speed Automated Driving (LSAD) systems. In addition, the proposed process is being considered for incorporation in an industry “recommended practice” document.

Appendices

TRUST CALIBRATION: CASE EXAMPLE[7]

Appendix A1

Literature Review

In chapter three, the concept of calibration of trust was introduced. Calibration of trust is defined as *"the process of adjusting trust to correspond to an objective measure of trustworthiness"* (Muir, 1994). It has been suggested in chapter three that calibration of trust is a multi-stage process (Figure 3.2):

- Stage A: Initial phase: Influenced by static knowledge and an internal mental model of the driver
- Stage B: Loss phase: Influenced by failures (frequency and severity). The rate of decrease is influenced by the consequences of these failures
- Stage C: Distrust phase.
- Stage D and stage E: Recovery phase.

[7] Contents of this chapter have been published in the following publication:

Khastgir, S., Birrell, S., Dhadyalla, G. and Jennings, P. (2017) 'Calibrating Trust to Increase the Use of Automated Systems in a Vehicle', in Stanton, N. et al. (eds) Advances in Human Aspects of Transportation. Advances in Intelligent Systems and Computing. Springer, Cham, pp. 535–546.

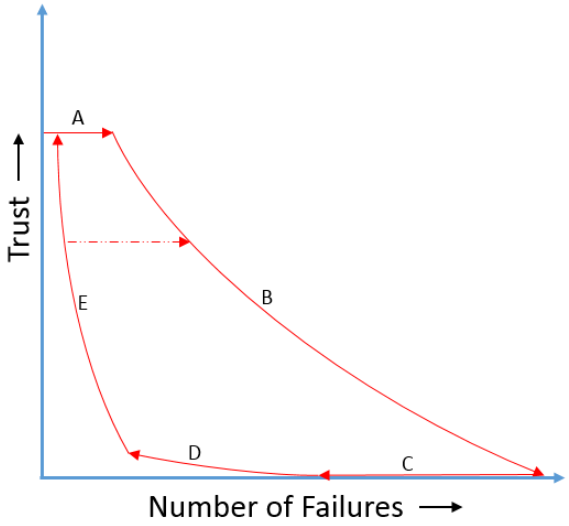

Figure A1.1: Trust calibration graph: Representing trust hysteresis

Chapter three also introduces the concept of *Trust Calibration Number* (TCN). TCN has three aspects: 1) Area enclosed by the trust calibration graph, referred to as Area of uncertainty 2) Slope of stage B 3) slope of stages C, D and E.

$$\begin{matrix} Trust \\ Calibration \\ Number \end{matrix} = \text{Area of uncertainty} + \text{Fcn}_1(\text{slope of stage B}) - \text{Fcn}_2(\text{slope of stages C,D, E})$$

$$\boldsymbol{Design\ Goal}: Minimize(Trust\ Calibration\ Number)$$

The area enclosed in the graph (Figure 3.2) is defined as the *area of uncertainty*. This chapter explains the calibration of trust using a case study of driver-take over in an automated vehicle.

A1.1. Take-over process with knowledge as intervention method

One of the open research questions in the area of driver-automation interaction is the take-over process which has been widely discussed in literature (Eriksson et al., 2017; Eriksson and Stanton, 2017a; Gold et al., 2013; Hergeth et al., 2017; Lorenz et al., 2014; Louw et al., 2016; Radlmayr et al., 2014). A take-over process is when the responsibility for the driving task shifts from the driver to the automated system or vice-versa. In this case study, the author discusses the latter. A take-over process is said to have various phases: 1) stable performance phase 2) event/condition state change phase or change in driver intent 3) take-over request phase 4) transfer phase and 5) receipt and recovery phase (Figure A1.2).

A take-over response is defined as *"the specific, measurable action taken by the human user or the system to partially or fully resume the Dynamic Driving Task (DDT)"* (SAE International, 2016a), where DDT is defined as *"All of the real-time operational and tactical functions required to operate a vehicle in on-road traffic, excluding the strategic functions such as trip scheduling and selection of destinations and waypoints, and including without limitation…(tactical)"* (SAE International, 2016b).

It is said that cognitive tasks as compared to motor tasks have more influence on take-over time (Zeeb et al., 2015). A study comparing the effect of different types of distraction on quality of take-over time showed that a cognitive distraction task resulted in a poorer quality of take-over as compared to motor distraction tasks (Zeeb et al., 2016). This emphasizes the need for increased SA2 for drivers to have better interaction with the automation and to react to take-over situations (Radlmayr et al., 2014). SA3 is needed to predict take-over situations which would result from a knowledge-based behaviour. While SA1 could mean eyes on the road, it doesn't guarantee SA2 (which requires cognition) (Radlmayr et al., 2014). Motor reactions could possibly occur as a result of intuitive actions due to drivers' skill-based behaviour, e.g. driving on the same route every day. However, the quality of a take-over manoeuvre depends on the driver's SA2 and SA3 which are a result of both rule-based and knowledge-based behaviour. Therefore, while considering takeover scenarios, it is important to consider both response times and the quality of take-over.

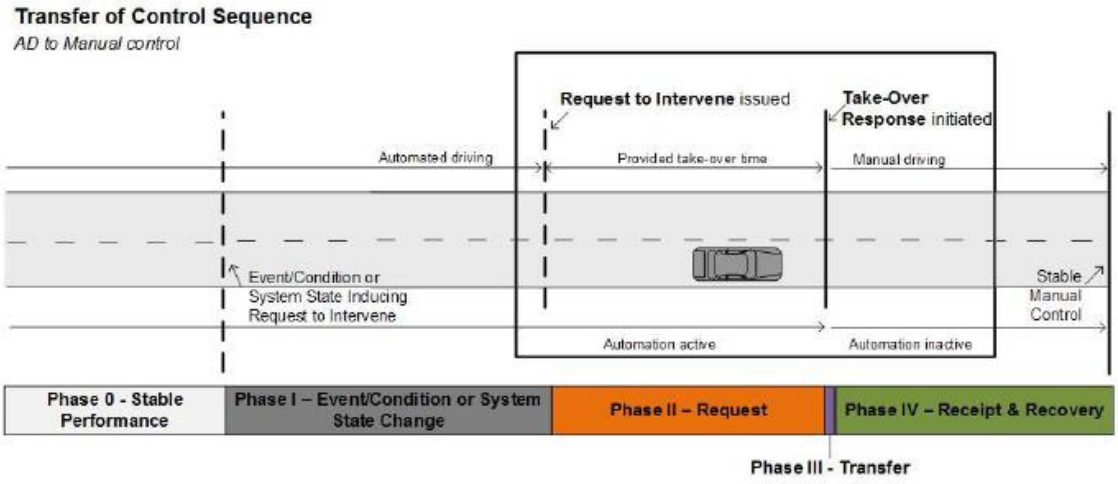

Figure A1.2: Different phases of take-over process (system-initiated). Adopted from SAE J3114 (SAE International, 2016a) *which has been adapted from* (Seppelt and Victor, 2016) *and* (Damböck et al., 2012).

Depending on whether the take-over process has been initiated by the automated system or the driver, the take-over request phase may or may not occur (Figure A1.3). For a particular

trip, the driver is responsible for making the strategic decision about when to initiate the AD mode or drive in manual mode. Automated Driving mode can perform both tactical and operational driving tasks. However, the decision to disengage AD mode based on contextual parameters is a tactical decision which the driver is supposed to make in order to ensure safe manoeuvring of the vehicle to achieve the strategic goals of the driving task. Irrespective of whether the take-over process has been initiated by the driver or the system, the driver has to perform the take-over response in order to take control back.

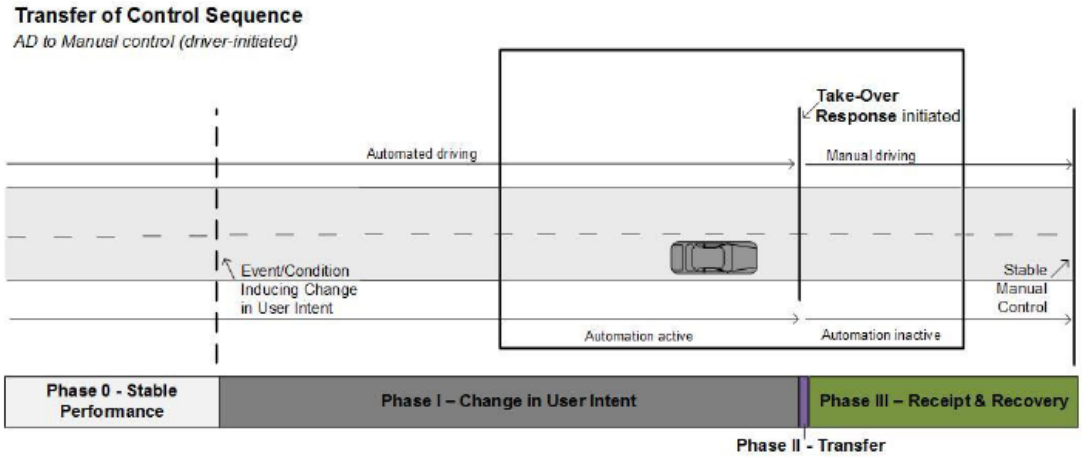

Figure A1.3: Different phases of take-over process (driver-initiated). Adopted from SAE J3114 (SAE International, 2016a) *which has been adapted from* (Seppelt and Victor, 2016) *and* (Damböck et al., 2012).

The quality of the take-over process is dependent on the driver's understanding of the state of the system and the environment (SA2) (Radlmayr et al., 2014). The quality of the take-over process would appreciably be better if the driver has a high level of SA3 and could predict a possible take-over process situation by making knowledge-based decisions (Rasmussen, 1983). Based on evidence from literature, the quality of take-over processes and their corresponding relationships was analysed with the calibration of trust. Similar to the stages of calibration of trust, Figure A1.4 has five points of interest (A-E). At point A, the automated system detects a situation to which it cannot respond appropriately or safely. At point B, urgent intervention is required from the driver in order to ensure the safety of the vehicle. At point C, the driver performs the intervention (take-over response). The driver tries to understand the current system state and decides on the appropriate intervention manoeuvre (e.g. applying the brakes, steering to an adjacent lane, applying acceleration etc.)

in the duration between points B and C. By the time the driver has reached point D, the driver has regained stable manual control of the vehicle.

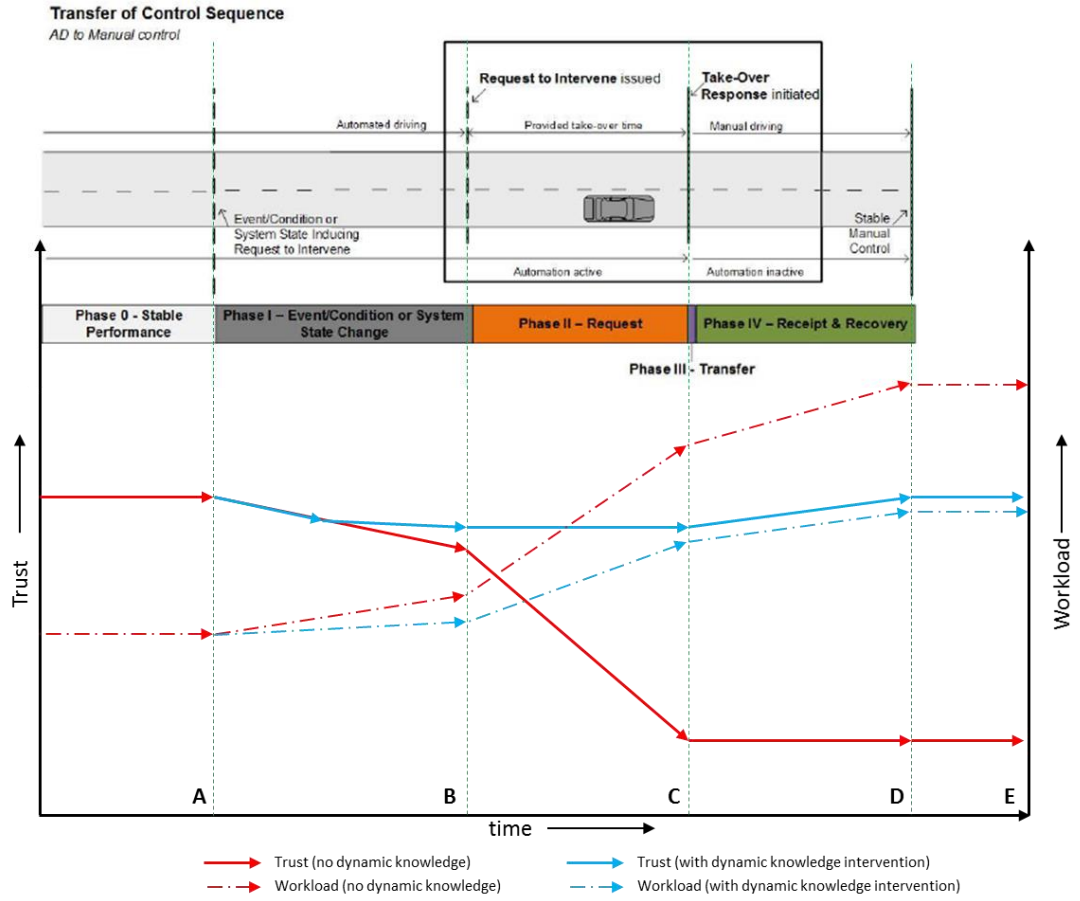

Figure A1.4: Calibration of trust with intervention methods in a take-over scenario (system-initiated)

The author will now discuss the described take-over process in two different contexts: 1) with no dynamic knowledge feedback provided by the automated system versus 2) with dynamic knowledge feedback provided by the automated system to the driver. In the absence of any dynamic knowledge (feedback) of the failure or a request for take-over, as the failure starts to occur at point A, trust on the system decreases gradually (Chavaillaz et al., 2016b) as it is not yet a critical situation, and the driver's workload increases. Due to highly critical nature of the stage between point B and point C, drivers' trust plummets to a low level as they find it hard to gain enough situational awareness to understand the required intervention and are caught by surprise. At the same time, drivers' workload increases rapidly due to the requirement of a high level of cognition. While the drivers' workload level stabilizes and decreases after point D (once stable manual control is achieved), drivers' trust level remains low as it is difficult for the drivers to recover trust quickly. However, the driving simulator studies (in chapter five) demonstrated that with the introduction of knowledge, trust levels can be calibrated to an appropriate level even with an imperfect

automation. They argued that introduction of (static) knowledge increased drivers' workload as it increased their cognitive workload. However, providing dynamic knowledge only or in conjunction with static knowledge overcame this effect, suggesting that drivers require real-time feedback on the automation system status in order to calibrate their trust.

TEST METHODS[8]

Appendix A2

Literature Review

Chapter six introduced three constituting elements of testing:

- Test scenarios
- Safety analysis
- Test methods

Test scenarios and safety analysis have been discussed in detail in chapter six – eight. This chapter provides an overview of the various test methods being used for Advanced Driver Assistance System (ADAS) and Automated Driving System (ADS). Furthermore, this chapter highlights some of the challenges in test methods for ADASs and ADSs, while introducing WMG's 3xD Simulator for Intelligent Vehicles which was created with a vision to answer some of these challenges.

[8] Contents of this chapter have been published in the following publications:

Khastgir, S., Birrell, S., Dhadyalla, G. and Jennings, P. (2015b) 'Identifying a gap in existing validation methodologies for intelligent automotive systems: Introducing the 3xD simulator', in Proc. of the IEEE Intelligent Vehicles Symposium 2015. Seoul, South Korea, pp. 648–653.

Khastgir, S., Birrell, S., Dhadyalla, G., Fulker, D., et al. (2015) 'A Drive-in, Driver-in-the-Loop Simulator for Testing Autonomous Vehicles', Proc. of the Driving Simulation Conference Europe 2015, pp. 117–122.

Khastgir, S., Birrell, S., Dhadyalla, G. and Jennings, P. (2015a) 'Development of a drive-in driver-in-the-loop fully immersive driving simulator for virtual validation of automotive systems', in Proc. of the IEEE Vehicular Technology Conference Spring 2015. Glasgow. doi: 10.1109/VTCSpring.2015.7145775.

Due to the safety critical nature of ADASs like collision avoidance, and lane keeping assist systems and ADS, testing methodologies which enable exhaustive testing need to be adopted for their validation to ensure their safe performance in all conditions. Moreover, these test methods require precise measurements, high reproducibility and test scenarios that are representative of the real world. However, the test scenarios used in the test methods include manoeuvres which are hard to reproduce, too expensive and potentially unsafe for human drivers (Hendriks et al., 2010).

Researchers have developed several innovative test methods (e.g. Coordinated Automated Driving, VEHiL, ViL, driving simulators etc. which have been discussed in sections 0 - 0) to meet the challenges to validate ADAS and other automated features present within cars. In ADASs, the driver is still in the loop of the driving task and up until SAE Level 3, can take over or asked to take over the control of the vehicle at any point in time. Thus, one of the biggest challenges faced by researchers and manufactures is to develop a test method that provides the ability to reproduce test scenarios and test results in a safe manner while having the driver-in-the-loop. The introduction of virtual testing has allowed researchers to increase their ability to reproduce results even in a driver-in-the-loop environment, however many challenges still exist. Some of test methods and the challenges associated with them are discussed in this chapter.

A2.1. On-Road testing

Traditionally, on-road testing involves deploying a nearly final version of the subject-under-test (SUT) into real-world traffic conditions. While traditional automotive systems require driver to be still responsible for ensuring safety of the driving task at all times, higher level ADAS and AD systems remove the driver from this responsibility. Additionally, due to the safety critical nature of ADAS and AD systems, they have the potential to endanger humans in case they don't function properly. The assumption behind on-road testing is that if the SUT has experience enough scenarios and is able to handle them in a safe manner, SUT has passed the tests and can be deployed. However, the challenge remains to understand the number of scenarios required in real-traffic to suggest that SUT has passed on-road testing. Until SUT has been proved to be safe, it requires a trained safety driver to be present behind the wheel at all times (Wachenfeld and Winner, 2017a). This makes on-road testing highly resource (cost and time) intensive. Therefore, it has been suggested that on-road testing just before the start of production for ADAS and AD systems will not be sufficient and suitable to establish their safety levels. This suggests the needs for other methods for testing ADAS and AD systems which may complement on-road testing.

A2.2. Coordinated Automated Driving

Certain test scenarios for driver assistance systems require two or more vehicles to be able to perform precisely synchronized manoeuvres, particularly with lateral dynamic manoeuvres. As human operation of vehicles is incapable of providing consistent results, and is also dangerous in accident-prone scenarios, a coordinated automated driving concept has been introduced by Daimler. In this method, the vehicles are driven by robotic systems controlling the steering and braking. The robots are programmed to follow the desired path and perform the required manoeuvres, thus allowing precise reproduction of the scenario. The robotic vehicles are equipped with an Inertial Measurement Unit (IMU) to measure the actual position. However, due to the presence of an inherent drift in an IMU, this system is backed up by a Differential Global Positioning System (DGPS) to ensure long term accuracy in the cm range (Schöner et al., 2009). The system allows for precise trajectory control of the vehicles as long as the manoeuvres stay away from the physical limits of vehicle dynamics.

The robotic vehicles move on an open test track. For safety critical scenarios they are driven in driverless mode. In other scenarios the driver has an option to switch between automatic control and human control with the press of a switch. The dynamic manoeuvres are of special interest to understand whether the system remains within existing legislative limits. As a safety system, an operator controls all vehicles via a WLAN network from a common base station. It also comprises of a controlled shut-down procedure to prevent any accidents due to sudden decelerations. An important feature of this method is the high degree of reproducibility even in scenarios requiring high deceleration rates. However, the setup cost is high as all vehicles need to be fitted with robotic systems and can only be done towards the end of product development.

A2.3. VEHiL: VEHicle Hardware-in-the-Loop

The VEhicle Hardware-in-the-Loop (VEHiL) concept introduced by TNO Automotive, the Netherlands, allows for evaluation of various driver assistance systems on full scale intelligent vehicles and their infrastructure in a laboratory (Verburg et al., 2002). First, a virtual environment (PRE-SCAN (Gietelink et al., 2004)) is defined in which the various interactions between vehicles and infrastructure are simulated. In a second step, the full scale Vehicle Under Test (VUT) is placed on a chassis dynamometer, which is fed with realistic road load and interfaced with the simulated environment. The benefit of this method is that the VUT is treated as a “black box” from which vehicle states are measured and fed into the simulated environment. The whole system works in a partly real and partly simulated environment combining the benefits of both types of testing. The added benefit of the

chassis dynamometer is that the system can accommodate a wide range of vehicle types, giving it added flexibility for validation.

It is interesting to note that in this method only the relative motion between the VUT and other traffic participants is reproduced. The other traffic participants are realized by a 4-wheel driven, 4 wheel steered robot vehicle (as shown in Figure A2.1) which responds to the motion commands (in terms of position, velocity and acceleration) of the traffic simulator and carries out the commands relative to the static VUT. This movement is sensed by the VUT sensors fooling it to believe it is actually moving on the road. Various types of ADAS systems can be tested using the VEHiL facility, however certain systems like the collision mitigation, lane keeping assist system and lane departure warning system cannot be tested to appropriate levels. Additionally, the setup of the full system requires a large open space and has considerable cost attached with it. On the positive side, various accident-prone scenarios can be validated and reproduced in this setup without harm to the human driver. Although, no level of virtual validation can completely replace in-field testing, VEHiL seems to provide a smooth transition from the full simulation world to in-field testing.

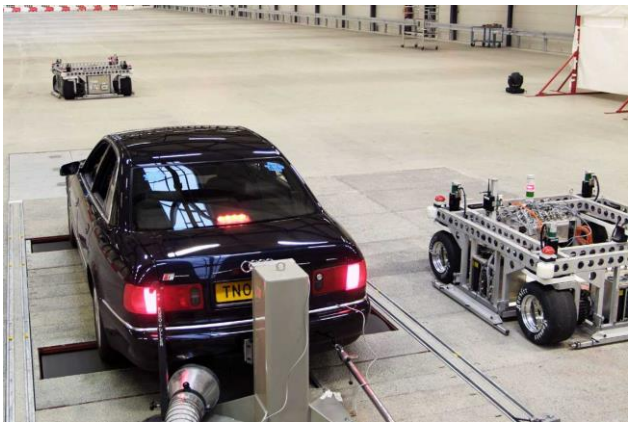

Figure A2.1: VEHiL

A2.4. ViL: Vehicle in the Loop

The concept of “Mixed Reality” which combines actual reality and virtual reality is another prospective approach for validation of ADAS (Blissing et al., 2013). Augmented Reality, a manifestation of mixed reality has been widely used in the ADAS validation technologies of late. Test tools and techniques described above do fulfil the requirements for safe, reproducible methods, but only within limits. For higher levels of automation, which include collision mitigation and lane keeping assist, further development of test methods was required. A host of driving simulators (which will be discussed in later sections) have been

made and test methodologies for ADAS systems have been developed. However, none of them have been able to provide a platform for validation of higher levels of automation in vehicles. The idea of the Vehicle in the Loop test setup introduced by Audi AG and later adopted by Volkswagen, combines a test vehicle and synthetic environment by means of Augmented Reality.

In vehicle in the loop configuration, the driver drives the vehicle in open spaces (a test track) without any traffic. Additionally, the driver wears a Head-Mounted Display (HMD) (as shown in Figure A2.2) which simulates the traffic and road environment in front of the driver by use of augmented reality. Since the whole traffic environment exists in the virtual world only, all safety critical scenarios can be easily validated with a high degree of accuracy and reproducibility. The vehicle's trunk houses two separate computers for data measurement and traffic flow simulation, along with the radar control unit. The vehicle's position is calculated by a three pronged approach of differential GPS, inertial navigation and couple navigation by integrative updating of the vehicle sensor data (Bock et al., 2005). This information along with the driver's head position which is monitored by a head tracker, is transferred to the traffic simulation software online.

The Field of View (FoV) of the driver whilst wearing the HMD is limited to the natural visual field which changes continuously according to the driver's head position. Since there is an interaction between the real world and the virtual world, real world sensors need to be simulated to enable the communication with the virtual world. Additionally, the ability to make manual triggers for various scenarios increases the flexibility of this system. The fidelity of the results is highly dependent on the fidelity of the modelled sensors and the realism of the augmented reality visuals. The scenario generation is done via a computer present in the trunk of the vehicle, which interacts with the HMD and the vehicle systems via CAN to generate the view for the driver. Another computer, also present inside the trunk, stores the captured data.

Figure A2.2: ViL

The setup of the vehicle in the loop system requires a large open space where the driver can drive the vehicle without any obstacles. Additionally, due to the presence of the HMD, this technique is not optimal for driver behaviour studies which form an essential aspect of the acceptance studies for higher level of autonomous features and ultimately a self-driving car. The degree of realism of the simulated environment is limited, as the augmented reality field is still under progress. In certain scenarios like a vehicle cut-in, the results are compromised (Bock et al., 2007).

A2.5. Driving Simulators

As mentioned in sections A2.1, even though on-road studies / real-world studies form a major part of testing of ADAS and ADS, they can be extremely resource (time and cost) intensive. Moreover, due to the safety critical nature of ADAS and ADS, it can be extremely difficult to perform certain driving manoeuvres which are evasive in nature in a safe and a reproducible manner. In chapter two, the need to better understand and evaluate the interaction between ADAS and ADS with drivers to develop mechanisms to increase their trust has been discussed. However, most of the on-road studies are undertaken with a trained safety driver behind the steering wheel. Thus, it is not possible to understand the driver-automation interaction as the trained safety drivers would behave differently than a normal drivers.

Driving simulators have been used as a possible tool to overcome some of the challenges of real-world testing and to provide a safe environment to test drivers' response to ADAS and AD functions. One of the major benefits of a driving simulator is that it offers a completely safe environment for validation of various technologies in a virtual world with the driver-in-the-loop and in a reproducible manner (Underwood et al., 2011). Ability to vary both the traffic and weather conditions in a reproducible manner, add more fidelity to the results from

a driving simulator study. In order to answer research question one and research objective two identified in chapter three on evaluating the effect of knowledge on trust, it was essential to use a setup that would enable the author to evaluate users' interaction with automated systems in a safe and reproducible manner. Thus, driving simulator was selected as the setup to answer research question one and meet research objective two. The studies discussed in this thesis were conducted on WMG's 3xD Simulator for Intelligent Vehicles (WMG, 2017).

A2.5.1. WMG 3xD Simulator for Intelligent Vehicles

For the driving simulator studies presented in this thesis, WMG's 3xD driving simulator was used (WMG, 2017). The 3xD simulator offers a unique platform for system validation, as well as for driver behaviour studies around autonomous vehicles with varied levels of automation. The WMG's 3xD simulator is housed in a radio-frequency (RF) shielded room with dimensions of 9.4 x 9.4 x 3.2 metres and has a 360° Field of View (FoV) which is achieved using eight projectors mounted from the ceiling. The 360° FoV is realised via a cylindrical screen which is eight metres in diameter and three metres in height. The dimensions allowed to have a Land Rover Evoque as the Built-Up Cab which was used for the studies. The 360° Field of View, the cylindrical screen and a life-size Land Rover Evoque as the ego vehicle (Figure A2.3), help immerse the participants in a driver-in-the-loop simulation environment.

Apart from the reasons for choosing a driving simulator (mentioned above), the choice of using the 3xD simulator was the ability to control the simulation (i.e., simulation entities like ego vehicle, other vehicles, environment, pedestrians etc.) in real-time via a TCP/IP Hardware-in-the-Loop (HiL) server-client interface. The real time control involves shifting the ego vehicle from manual driving mode to automated driving mode, triggering events in the simulation environment, applying emergency braking of the ego vehicle etc. The HiL server-client interface allowed external microprocessor to be connected to the simulator. For the studies one, two and six, a Raspberry Pi board was connected to the simulator via the HiL interface (Figure A2.4). A Raspberry Pi board was chosen as it offered the ability to connect the inputs of an analogue switch (button) (via the GPIO pins on the Raspberry Pi) to stimulate the 3xD simulator.

Figure A2.3: WMG, 3xD Simulator for Intelligent Vehicles

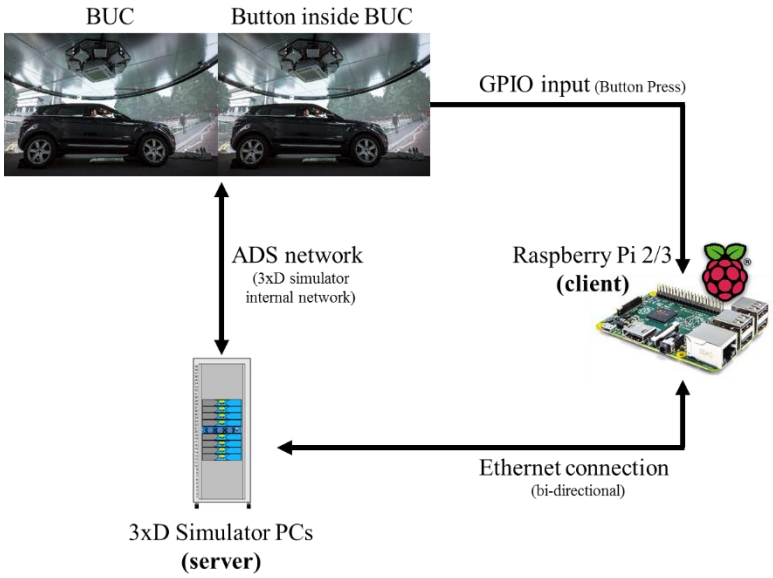

Figure A2.4: 3xD simulator interface with Raspberry Pi via the HiL server-client interface

The challenges offered in the setup of the 3xD simulator were due to the mechanical concept and the usage envisioned for it. Some of the research challenges included: steering wheel feedback, RF immersive environment, motion cueing and degree of realism.

RF Immersive Environment

The simulator itself is housed inside a "Faraday cage" with dimensions 9.4 x 9.4 x 3.2 metres. The faraday cage provides Radio Frequency (RF) shielding to the setup in simulator room. This is required as the 3xD simulator has the ability to artificially create an immersive RF environment which includes generation of GPS and wireless communication signals such

as the IEEE 802.11, LTE, 5G etc. It is important to note that the 3xD simulator wasn't envisioned as a vehicle dynamics evaluation platform, therefore motion cueing inputs for the driver are limited. However some degree of pitch, roll and vibration inputs are provided to the driver by a fixed motion base and actuator setup to provide an immersive environment to the driver.

30 mile LiDAR route

In order to increase the immersivity of the driver during the simulation, it is important to present a real world scenario to the driver. In order to have a real-world base map, a 30 mile geo-specific database was commissioned around Coventry, UK, which includes the main ring road, a section of the motorways (M6 and M42) and a number of interconnecting A-roads, as shown in Figure A2.5. The route was scanned using a LiDAR (Laser Illuminated Detection And Ranging) sensor system which provided a road terrain accuracy of up to 2cm.

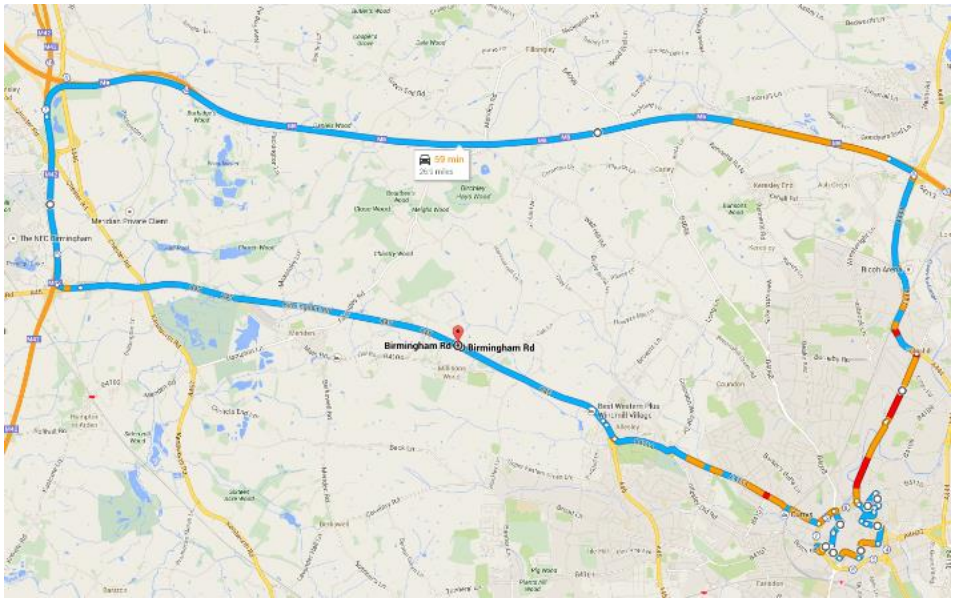

Figure A2.5: 3xD simulator LiDAR scanned route

Steering Wheel feedback

The drive-in feature required the ability to provide steering feedback to the driver with the ignition turned off. As is commonly known, the power steering system is extremely stiff with the engine off as the power assisted steering isn't available. On a typical simulator with a modified BUC the steering wheel is disconnected from the road wheels and an electric motor system is connected to the steering column to allow centring and provide road cues, which is not possible on a vehicle that is driven in and can't be modified.

The mechanism developed for the drive-in system consists of two separate steering plates that the front wheels are placed on. The main plate consists of a circular rotatable plate to which one of the wheels is mounted firmly. The plate is then driven by a standard steering feedback motor, as would normally be connected to a traditional simulator BUC, with an appropriate gear ratio to provide the centring and feedback. The other plate is much simpler with a low friction bearing that allows lateral movement. Figure A2.6: Active steering plate for steering feedback shows the wheel of the vehicle mounted on the active steering plate.

The risk of making both the plates active would lead to a possibility of them not being 100% synchronized and the motors not being able to deliver the desired steer. Additionally, to make the motor drive work, the wheel needs to be held accurately at the correct position to ensure it rotates around the correct part of the plate. If this was done for both the wheels it would very difficult to get both plates correctly positioned, each time a new vehicle is driven in. The secondary passive plate has an ability to not only rotate but also slide in both lateral and longitudinal directions.

Figure A2.6: Active steering plate for steering feedback

Motion Cueing

The requirement to have some form of motion cueing for the BUC needed the addition of actuators to the BUC vehicle itself. Systems typically use floor mounted actuators to move each corner of the vehicle up and down. However in this case there where two issues. Firstly due to the drive-in vehicle, there was the need to move the BUC from the simulator to allow it to be replaced. This meant that the actuators needed to be removed separately from the vehicle and then carefully aligned during replacement, making it a difficult and time consuming process. The second issue was the weight of the selected BUC donor vehicle, a

Land Rover Evoque, which even after the removal of non-essential elements had a weight of approximately 300Kgs at each corner. This was at the limit of the available off the shelf actuator systems. The actuators chosen have a stroke length of 150 mm which has been mechanically reduced to 100 mm (due to ergonomic issues in the BUC). They have a maximum frequency of 100Hz providing a maximum pitch of ± 2.1° with a maximum roll of ± 4.8°.

The original plan was to mount the actuators in place of the shock absorbers but due to the weight issue and the length of the actuators the solution shown in Figure A2.6 was adopted. It involves the conversion of linear horizontal motion of the actuators into vertical motion via a specially design lever mechanism. The system is mounted under the bonnet and in place of the rear axle. This mechanism provides the ability to control each of the actuators independently.

Figure A2.7: Actuator assembly of the BUC

This has a number of benefits over the original solution:

- It gives about a 1/3 extra lift but with a reduced movement capability by means of the levers.
- It also removes the need to modify the suspension system to accommodate the length of the actuator system

Table A2.1: Specifications for 3xD Simulator

Features	**Specification**
Max. Payload	3000 kg
Screen diameter	8 m
Screen Height	3 m
Number of projectors	8
Field of view	360°
Max. BUC roll	± 4.8°
Max. BUC pitch	± 2.1°

Table A2.2 shows a comparison between driving simulators and other test methods discussed earlier.

Table A2.2: Comparison of various test methods and WMG 3xD simulator

Features	**Coordinated Driving**	**VEHiL**	**Vehicle in the Loop (ViL)**	**WMG 3xD Simulator**
Aim of methodology	Testing ADAS systems	Testing ADAS systems	Testing ADAS systems	Testing various levels of automation in vehicles
Driver-in-the-Loop	No (in critical scenarios)	Yes (Less degree of realism)	Yes (with Head Mounted Display)	Yes
Vehicle Interchangeability	Yes but all systems need to be re-fitted	Yes	Yes	Yes
Intelligent Features that can be system validated	ACC, Collision Mitigation, LKAS (Limited ADAS)	ACC, Collision Mitigation, Fault Injection (Limited ADAS)	ACC, Collision Mitigation, LKAS (Limited ADAS)	ACC, CACC, Collision Mitigation, LKAS (ADAS), Level 3-4-5 automation, Path Planning Algorithms
Probability of Human Accidents	High (when driver is in the loop)	Low	Low	Zero (Since vehicle ignition is off)
Driver Field of View (FoV)	Not Applicable (no driver)	Full	Limited	Full 360°
Fidelity of Vehicle Dynamics (on a scale of 0-10)	Not Applicable (no driver)	3	8	4
Possibility of Driver Behaviour studies	No	No	Yes but not optimal	Yes
System Under Test	Modified Vehicle	Vehicle	Vehicle	Vehicle and BUC
Relative Cost (on a scale of 0-10)	5	8	4	4

SPEED AND CONTROLLABILITY: A DRIVING SIMULATOR STUDY

Appendix 3

As discussed in chapter 6, the ISO 26262 – 2018 standard is industry's gold standard for functional safety. The standard refers to ASIL as a metric for analysing the risk. An ASIL rating comprises of three components: Severity (S), Exposure (E) and Controllability (C). However, the standard fails to provide objective guidance to evaluate each of the three (S, E and C) ratings or identify the parameters influencing these ratings (Yu et al., 2016). The lack of objectivity leads to subjective interpretation of the standard by experts causing intra-rater and inter-rater variation leading to reduced reliability of the ratings (Ergai et al., 2016).

The ISO 26262-2018 (ISO, 2018c) standard states that *"The evaluation of the controllability is an estimate of the probability that someone is able to gain sufficient control of the hazardous event, such that they are able to avoid the specific harm"*. The standard classifies controllability into four classes: C0 – C3. However, it doesn't provide an objective manner of assigning this rating and thus, introducing subjectivity to the risk rating classification. As discussed earlier in chapter five, the risk rating forms a part of the dynamic knowledge and unreliable knowledge can lead to low trust.

Thus, to answer research question three focussed on improving the reliability of the knowledge, the concept of objectivising severity and controllability ratings was introduced in chapter eight.

One of the major factors influencing controllability of the hazardous situation is the speed of the vehicle. In order to evaluate the effect of speed on controllability of a vehicle, a driving simulator experiment was done. This chapter discusses the effect of speed on controllability

of the driver in a low-speed automated driving system through a driving simulator study. Subsequently, an initial controllability rule-set was developed.

A3.1. Study Method

A3.1.1. Participants

Similar to other driving simulator experiments (discussed in chapter 5), ethical approval for the experiment was secured from the University of Warwick's Biomedical & Scientific Research Ethics Committee (BSREC) (REGO-2015-1746 AM02). Forty four participants (11 female and 33 male) were recruited via email invitations. Participants were at least 21 years of age and were required to have a valid driving license. Participants experienced a maximum of three study runs (each at a different speed), one at low speed (5 – 10 miles per hour: band A), one at medium speed (14 – 16 miles per hour: band B) and one at high speed (20 – 25 miles per hour: band C). There were six possible permutations for the order (${}^{3}P_{2}$) in which participants would experience the three speeds. The order in which participants experienced various speed bands was randomized.

A3.1.2. Study Design

The experiment was designed as a 1 x 3 factorial design with vehicle speed band as a within subject variable. In band A (5 – 10 mph) and band C (20 -25 mph), there were six speeds and in band B there were three speeds (14 – 16 mph). Each band had 36 data points from 36 individual runs. While most participants completed three runs, some participants had to drop-out in between runs due to the onset of simulator sickness. In case a participant experienced simulator sickness during a run, that run was removed from the data analysis and the study for the participant was stopped. However, any previous study run that the participant might have successfully completed was taken into consideration for data analysis.

As part of the study, each participant was driven in automated mode in all three runs and participants experienced 19 hazardous situations in each run. All hazardous situations required driver intervention to prevent an accident. In order to intervene, participants were asked to press an emergency stop button provided to them. Pressing the emergency stop button applied full braking power of the vehicle and the vehicle eventually came to a stop. Participants were seated in the front passenger's seat, therefore didn't have access to brake pedal or steering wheel. In case the participants met with an accident, the simulation was first paused, participants were told that they have met with an accident and then the simulation was restarted from a safe state from a point preceding to the accident impact

point. Each complete run lasted around 10 - 15 minutes (depending on the speed band). The driving simulator route was a built-up world with bends, roundabouts, straight roads and T-junctions. The route had urban and rural sections. As the study run was in urban and rural sections, all hazardous situations were at low speed (less than 25 m/hr). Hazardous situations involved interactions with pedestrians (adults and children), cyclists, vehicles and motorcyclists.

Similar to study one (trust and static knowledge) (in chapter 5), this study also had a gamification aspect to it (Table A3.1). Participants were told that they will receive 1 point for every second spent in automated mode. In case they applied the emergency brake to evade a hazardous situation, they received a bonus of 200 points. However, a wrong application of emergency brake led to a penalty of 200 points. In case of a crash, participants lost 10000 points. Participants' goal was to maximise their score.

Table A3.1: Scoring criteria for study (gamification)

Type of Action	Points
Automated mode	1 / second
Correct Stoppage of the automated vehicle	+200
Incorrect Stoppage of the automated vehicle	-200
Crash	-10000

A3.1.3. Study procedure

Similar to driving simulator studies discussed earlier (in chapter 5), when participants arrived for the experiment, they were initially briefed about the experiment and informed consent was taken. In order to familiarise the participants with the driving simulation environment, each participant was given a trial run on the driving simulator which lasted 5 minutes. Participants were told that they can ask for as many trial runs as they wish in order to make themselves comfortable with the simulator environment. After the trial runs, participants were asked if they wished to continue with the study. In case the participants agreed, they experienced three runs at three different speeds (corresponding to the three different speed bands in a random order). At the end of each run, participants were asked to fill a simulator sickness questionnaire (SSQ) (Kennedy et al., 1993) in order to judge the onset of simulator sickness.

A3.2. Results

A3.2.1. Accidents

During the simulation run, in case the participant's vehicle (ego vehicle) collided with a pedestrian, cyclist or a vehicle, the simulation came to a stop (frozen state) and the instance was classified as an accident. As mentioned in section A3.1.1, participants experienced 19 hazardous situations which could lead to 19 potential accidents. The average number of accidents experienced by participants in the three speed bands has been illustrated in Figure A3.1. A one-way ANOVA was conducted for the average number of accidents with speed bands as a between group variable. Speed bands had a significant effect on the number of accidents, $F(2, 105) = 14.737$, $p < 0.05$. Post-hoc test revealed that while average number of accidents was not significantly different between band A and band B ($p > 0.05$), the average number of accidents in band C was significantly higher than band A ($p < 0.001$) and band B ($p < 0.001$). While the higher number of accidents in band C could potentially be due to the higher speed leading to lower reaction time, the similar ratings for band A (low speed) and band B (medium speed) were an interesting result which will be discussed in detail in section A3.3.

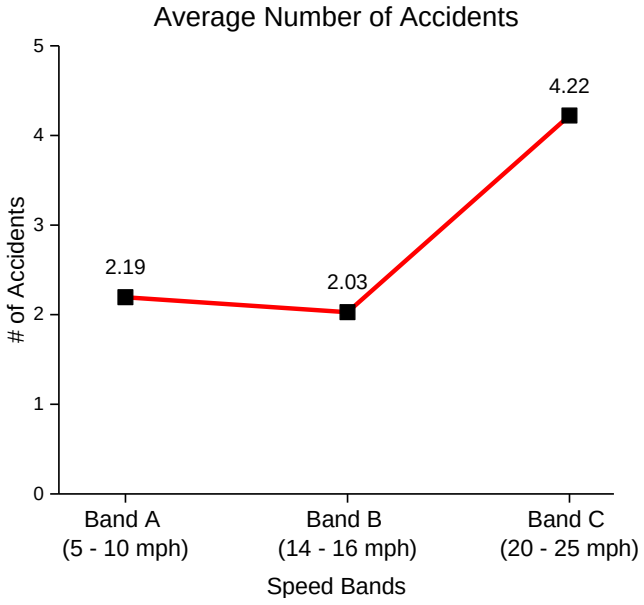

Figure A3.1: Average number of accidents in each speed band

A3.2.2. Missed presses

The average number of missed presses for the three speed bands showed an opposite trend as compared to the average number of accidents. The average number of missed presses

experienced by participants in the three speed bands has been illustrated in Figure A3.2. A one-way ANOVA for the average number of missed presses with speed bands as a between group variable suggested a significant difference in the average number of missed presses between the speed bands, $F(2,105) = , p < 0.001$. Post-hoc analysis revealed that the average number of missed presses for band A was significantly higher ($p < 0.001$) at 0.86 as compared to 0.167 for band B and 0.083 for band C. This is another interesting result which will be discussed in detail in section A3.3. There was no significant difference ($p > 0.05$) between the average number of missed presses for band B and band C.

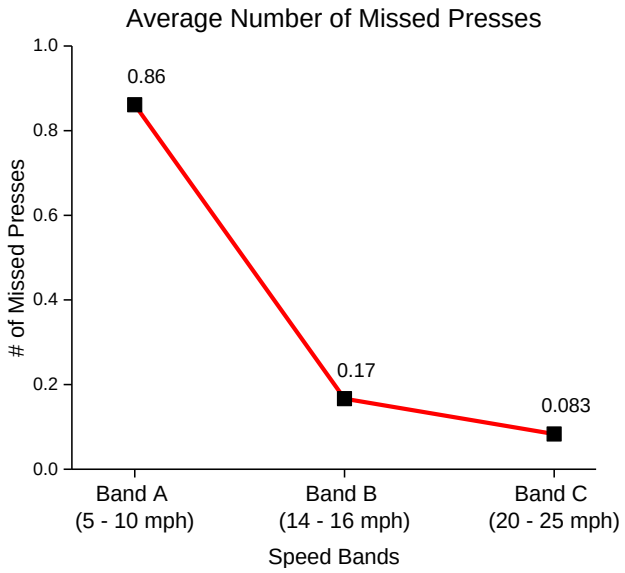

Figure A3.2: Average number of missed presses

A3.2.3. False presses

The average number of false presses experienced by participants in the three speed bands has been illustrated in Figure A3.3: Average number false presses. The average number of false presses for band A was 0.583, for band B was 0.917 and for band C was 0.833. While a visual difference could be observed in Figure A3.3 between average numbers of false presses for the three bands, the difference was not statistically significant ($p > 0.05$).

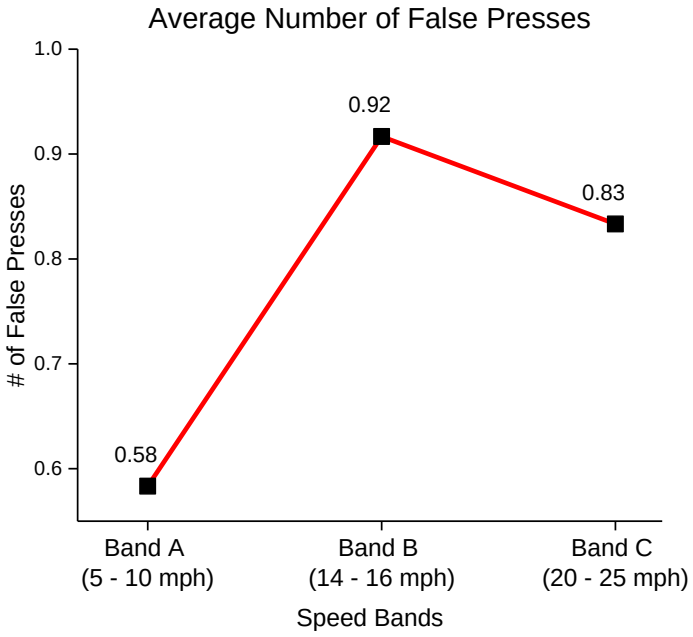

Figure A3.3: Average number false presses

A3.2.4. Speed Controllability Number (SCN)

Based on the number of accidents, false presses and missed presses, the author created a new term, Speed Controllability Number (SCN), to encapsulate the contents into one, to give a broader understanding of the ability of the participants to control the hazardous situation and prevent an accident. SCN captures within itself, aspects of mistrust and distrust (discussed in chapter 3 and chapter 5) and their effect on controllability. The calculation of SCN is described in equation 1.

$$\text{Speed Controllability Number (S.C.N.)} = 100 \times \left(\frac{\text{No. of Correct button presses}}{\text{No. of Required button presses}} - \frac{\text{No. of crashes when button pressed}}{\text{No. of Required button presses}} - 1.5 \times \frac{\text{No. of crashes when no button pressed}}{\text{No. of Required button presses}} - 0.5 \times \frac{\text{No. of false presses}}{\text{No. of Required button presses}} \right) \quad \ldots (1)$$

The first term in SCN represents *"appropriate trust"* (discussed in chapter 3 and chapter 5) and the participant's ability to react correctly by providing an active intervention (by pressing the emergency stop button) to stop the vehicle in response to a hazardous situation. Second term represents late realisation by the participants about the inability of the automated system to react to the hazardous situation. Third term represents high over-trust

leading to no intervention to a hazardous situation. Fourth term represents distrust in the automation system, and participant's intervention, even though no hazardous situation was present.

The average SCN in the three speed bands has been illustrated in Figure A3.4. The average SCN for band A was 79.8, band B was 86.5 and for band C was 74.2. A one-way ANOVA for SCN values with speed bands as a between groups variable suggested a significant difference between the groups, $F(2, 105) = 6.502$, $p < 0.05$. A post-hoc test revealed that the average SCN for band B was significantly higher than SCN for band C ($p < 0.001$) and nearly significantly higher than SCN for band A ($p = 0.053$). The SCN for band A and band C was not significantly different ($p > 0.05$).

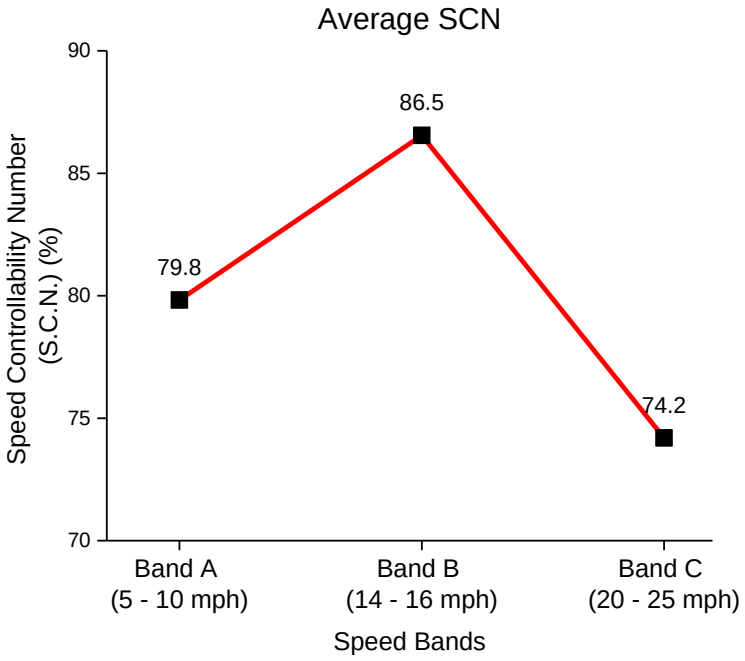

Figure A3.4: Average Speed Controllability Number (SCN)

A3.3. Discussion

Conventional logic would have suggested that the number of accidents would be highest at high speeds and lowest at low-speeds. Higher number of accidents at high speeds could be attributed to the lack of time to react to a hazardous situation at higher speeds. While the results of the study do suggest higher average number of accidents for band C (high speed), the lack of difference in the average number of accidents for band A (low-speed) and band B (medium speed) is an interesting finding. This finding can be better understood with the

discussion of the trend observed for average number of missed presses. The average number of missed presses was significantly higher for band A as compared to band B and band C. It is worthwhile to note that a missed press would lead to an accident. At low-speeds while participants had more time to react to a hazardous situation, they had higher inherent trust in the automated system's ability to react to the situation, suggesting over-trust on their part. The hypothesis is further corroborated by some of the qualitative feedback provided by the participants. Participant # 9 commented: *"false sense of security (at low speed). Harder to stay attentive at this speed"*. Another participant commented: *"I let the system do what it could. I expected at low speeds it should be able to handle"*.

The higher SCN value for band B as compared to band C and band A is an interesting finding suggesting the existence of an optimum speed range for low-speed automated driving systems, which is neither too low (giving a false sense of security), nor too high (leaving little time to react for the driver/operator). However, like any other driving simulator studies, the results from this study would need to be validated in a real-world environment.

A3.3.1. Initial Controllability rule-set

The findings of this study were incorporated in the controllability rating rule-set. The controllability parameters were mainly influenced by the vehicle's ability to change trajectory and the environment affecting vehicle's ability to make this change (McGehee et al., 2000; Rosén et al., 2011; Schaap et al., 2008; Young and Stanton, 2007). The parameters identified for controllability were: 1) vehicle velocity 2) time-to-collision (TTC) 3) distance to obstacle 3) maximum acceleration/deceleration 4) availability of safe area 5) road friction 6) gradient of slope. Time-to-collision (TTC) is defined as *"the time taken by the trailing vehicle to crash into the front vehicle, if the vehicles continue in the same path without adjusting their speeds"* (Chin and Quek, 1997). A condensed version of the controllability rule-set was used in the pilot study due to logistical reasons and is depicted in Table A3.2.

Table A3.2: Initial Controllability rule-set

Emergency Deceleration Value	Distance to Obstacle	TTC	Vehicle velocity	Controllability Rating
0.4g - 0.8g	< 6 m	< 1.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C0
			> 16 km/h	C2
	> 6 m	< 1.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C2
		1.0 - 2.0 sec	< 11 km/h	C0
			11 - 16 km/h	C0
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C0
			> 16 km/h	C1

Emergency Deceleration Value	Distance to Obstacle	TTC	Vehicle velocity	Controllability Rating
< 0.4g	< 6 m	< 1.0 sec	< 11 km/h	C3
			11 - 16 km/h	C2
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C2
			> 16 km/h	C3
		> 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C3
	> 6 m	< 1.0 sec	< 11 km/h	C3
			11 - 16 km/h	C2
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C3
		> 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C2

A3.4. Conclusion

The results from the driving simulator experiment suggest that there exists an optimum speed at which controllability of the vehicle is maximum. This optimum speed lies in the medium speed band. While no conclusion could be made whether low speed band has higher controllability or high speed band, it could be conclusively suggested that both (low and higher) have lower controllability as compared to medium speed band. This is an interesting finding, albeit counter-intuitive.

OBJECTIFICATION OF HARA WORKSHOP 2 (SWEDEN)

Appendix 4

A4.1. Workshop 2 (Sweden)

A4.1.1. Participants

Seventeen participants were recruited for the workshop conducted in Sweden. Participants had prior experience in functional safety assessment and most of the participants were involved in the Swedish functional safety technical committee involved in formulating Swedish comments for the ISO functional safety draft standards. Participants were grouped into five groups. Three groups had three participants while two groups had four participants. All participants were from Sweden.

A4.1.2. Workshop structure

The workshop consisted of an introduction that was followed by four rounds of 40 minutes each. Like in workshop one, each group was provided with two different hazardous events and were asked to rate the two given hazardous events. The same hazardous events were given in each of the four rounds. Figure A4.1 shows the workshop structure.

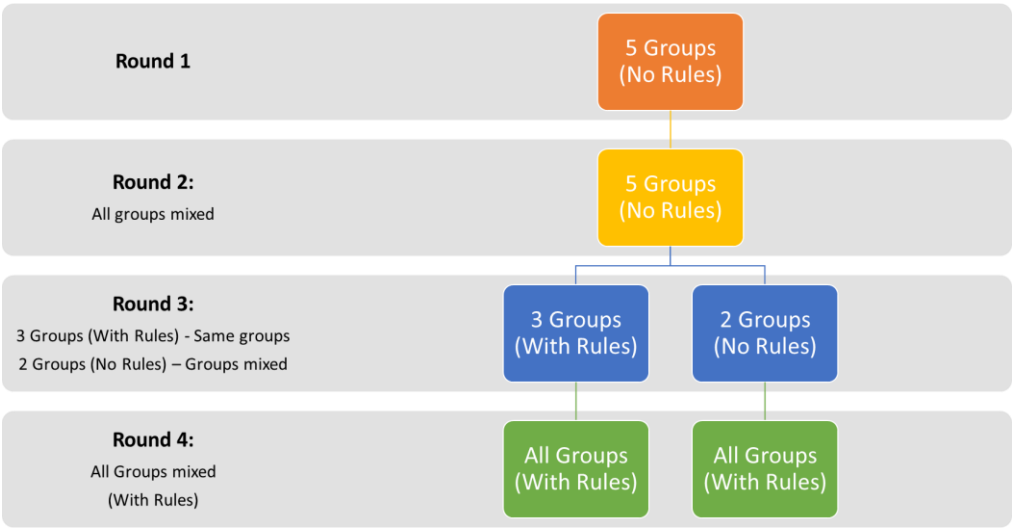

Figure A4.1: Sweden workshop structure (workshop 2)

A4.1.2.1. Additional System definition

In addition, to the system definition for the SAE Level 4 Low-Speed Automated Driving system (LSAD) described in chapter 8, participants were also provided with the following description:

- Vehicle crashworthy up to top speed
- Driving domain: public roads in inner city in UK

The additional information was provided as it was received as a part of the feedback from workshop 1.

A4.1.2.2. Hazard and hazardous event definition

The hazard provided to the participants was *"Unintended stopping or accelerating to a vulnerable position resulting in a collision"*. Based on the hazard, participants were provided two hazardous events and were asked to discuss the HARA for the two given events to give Severity, Exposure and Controllability ratings. The two hazardous events provided to the participants were:

- Pod travels into pedestrian
- Pod does unintended braking

A4.1.3. Rule-set

Based on the learnings from workshop 1 (discussed in chapter eight), a rule-set for exposure rating was introduced for workshop 2. In order to provide a better understanding on how to use the rule-set, through an example hazardous event, participants were explained in detail how to apply the rule-set. Additionally, with a modified definition for the hazard and by providing the context to the hazard, more information was the hazardous events was provided to the workshop participants.

A4.1.3.1. Severity rule-set

Severity rule-set has been depicted in Table A4.1 (same as the one used in workshop 1).

Table A4.1: Severity rule-set (workshop 2)

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Severity Rating
Pedestrian	< 11 km/h	< 2 km/h	S0
		< 6 km/h	S1
		< 12km/h	S1
	11 - 16 km/h	< 2 km/h	S1
		< 6 km/h	S2
		< 12km/h	S2
	> 16 km/h	< 2 km/h	S2
		< 6 km/h	S3
		< 12km/h	S3

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Severity Rating
Infrastructure	< 11 km/h	0 km/h	S0
		0 km/h	
		0 km/h	
	11 - 16 km/h	0 km/h	S1
		0 km/h	
		0 km/h	
	> 16 km/h	0 km/h	S2
		0 km/h	
		0 km/h	

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Severity Rating
Vehicle	< 11 km/h	< 10 km/h	S0
		< 20km/h	S1
		> 20 km/h	S2
	11 - 16 km/h	< 10 km/h	S1
		< 20km/h	S1
		> 20 km/h	S2
	> 16 km/h	< 10 km/h	S1
		< 20km/h	S2
		> 20 km/h	S3

Type of Obstacle	Vehicle Velocity	Oncoming Obj. Velocity	Severity Rating
Cyclist	< 11 km/h	< 8 km/h	S0
		< 14km/h	S1
		< 20km/h	S2
	11 - 16 km/h	< 8 km/h	S1
		< 14km/h	S2
		< 20km/h	S2
	> 16 km/h	< 8 km/h	S2
		< 14km/h	S2
		< 20km/h	S3

A4.1.3.2. Controllability rule-set

Controllability rule-set has been depicted in Table A4.2 and Table A4.3. In addition to the rule-set used in workshop 1, an additional rule-set was created for acceleration.

Table A4.2: Controllability rule-set - part 1 (workshop 2)

Emergency Deceleration Value	Distance to Obstacle	TTC	Vehicle velocity	Controllability Rating
< 0.4g	< 6 m	< 1.0 sec	< 11 km/h	C1
			11 - 16 km/h	C2
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C2
			> 16 km/h	C3
		> 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C3
	> 6 m	< 1.0 sec	< 11 km/h	C3
			11 - 16 km/h	C2
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C3
		> 2.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C2
Emergency Deceleration Value	**Distance to Obstacle**	**TTC**	**Vehicle velocity**	**Controllability Rating**
0.4g - 0.8g	< 6 m	< 1.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C0
			> 16 km/h	C2
	> 6 m	< 1.0 sec	< 11 km/h	C2
			11 - 16 km/h	C1
			> 16 km/h	C2
		1.0 - 2.0 sec	< 11 km/h	C0
			11 - 16 km/h	C0
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C0
			> 16 km/h	C1

Table A4.3: Controllability rule-set - part 2 (workshop 2)

Acceleration Value	Distance to Obstacle	TTC	Vehicle velocity	Controllability Rating
< 0.1g	< 6 m	< 1.0 sec	< 11 km/h	C1
			11 - 16 km/h	C2
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C0
			11 - 16 km/h	C1
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C0
			11 - 16 km/h	C0
			> 16 km/h	C1
	> 6 m	< 1.0 sec	< 11 km/h	C0
			11 - 16 km/h	C1
			> 16 km/h	C2
		1.0 - 2.0 sec	< 11 km/h	C0
			11 - 16 km/h	C0
			> 16 km/h	C1
		> 2.0 sec	< 11 km/h	C0
			11 - 16 km/h	C0
			> 16 km/h	C0
Acceleration Value	**Distance to Obstacle**	**TTC**	**Vehicle velocity**	**Controllability Rating**
0.1g - 0.4g	< 6 m	< 1.0 sec	< 11 km/h	C2
			11 - 16 km/h	C2
			> 16 km/h	C3
		1.0 - 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C2
	> 6 m	< 1.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C2
		1.0 - 2.0 sec	< 11 km/h	C1
			11 - 16 km/h	C1
			> 16 km/h	C2
		> 2.0 sec	< 11 km/h	C0
			11 - 16 km/h	C0
			> 16 km/h	C1

A4.1.3.3. Exposure rule-set

As per the learnings from workshop 1, a rule-set for exposure ratings was created and it was provided to the participants in workshop 2. Exposure rule-set has been depicted in Table A4.4.

Table A4.4: Exposure rating rule-set (workshop 2)

Area	Driving Domain	Country	Exposure rating
City Centre	Pedestrian Pathways	India	E4
		Sweden	E1
		UK	E2
		Germany	E0
	Normal road	India	E4
		Sweden	E1
		UK	E2
		Germany	E1

Area	Driving Domain	Country	Exposure rating
Sub-urban areas	Pedestrian Pathways	India	E3
		Sweden	E0
		UK	E1
		Germany	E0
	Normal road	India	E4
		Sweden	E0
		UK	E1
		Germany	E0

A4.1.4. Results

A4.1.4.1. Quantitative results

Just like workshop 1, each of the groups were asked to provide ratings for all three ASIL components, Severity, Exposure and Controllability. The exposure rule-set was introduced based on the learnings from workshop 1. The ASIL ratings given in various rounds to the two hazardous events have been depicted in Figure A4.2. The rounds have been plotted in x-axis and ASIL levels (QM – ASIL D) on the y-axis. In the first round (when no rules were provided), for hazardous event 1, the five groups gave four different ratings ASIL ratings, and three different ASIL ratings for hazardous event 2. The difference between the highest and lowest rating for both the hazardous events was of the order of three (ASIL D – ASILA). After round one and before round two, participants in the groups were shuffled. In round two, a similar variation in ASIL ratings was observed. For hazardous event 1, the five

groups gave three different ratings and for hazardous event 2, the five groups gave four different ratings. The difference between the highest and lowest ratings for both the hazardous events was again of the order of three (ASIL D – ASIL A for hazardous event 1 and ASIL C – QM for hazardous event 2). It is interesting to note that each of the groups provided a different justification (based on their assumptions) for their ratings. Once again, the wide variation in ASIL ratings illustrate the existence of inter-rater variation in the automotive HARA process, even when done by experts. The variation HARA ratings and experts' justification will be discussed in more detail in the qualitative analysis section (0).

Before round three, three of the groups (groups 1, 2 and 3: same groups as round two) were provided with ruleset for HARA. No ruleset was provided to the remaining two groups (group 4 and group 5). These two groups were mixed among each other for round three. For hazardous event 1, while groups 1, 2 and 3 used the rule-set, there was still variation in the ratings provided by them. While group 2 and 3 agreed in their rating (ASIL B), group 1's rating differed by an order of one (ASIL A). The ratings provided by the two other groups (group 4 and group 5) differed from the ratings of the three groups who used the rule-set by an order of one and two. For hazardous event 2, a similar trend was observed. Groups 1, 2 and 3, even with using the ruleset, differed in their ratings by an order of one (ASIL A – QM). While the two other groups (group 4 and 5), differed in their rating by an order of two (ASIL C – ASIL D).

In round four, all five groups were mixed and were told to use the rule-set to perform the HARA to give their ratings for the hazardous events. For hazardous event 1, the five groups gave three different ASIL ratings, with a maximum difference of the order of two (ASIL B – QM). For hazardous event 2, the five groups gave two different ratings, with a difference of the order of one between those ratings (ASIL A – QM). The ASIL ratings with the introduction of the ruleset show a visual decrease in variation in Figure A4.2. However, a deeper analysis of the ASIL components (Severity (S), Exposure (E) and Controllability (C)) provides an interesting insight.

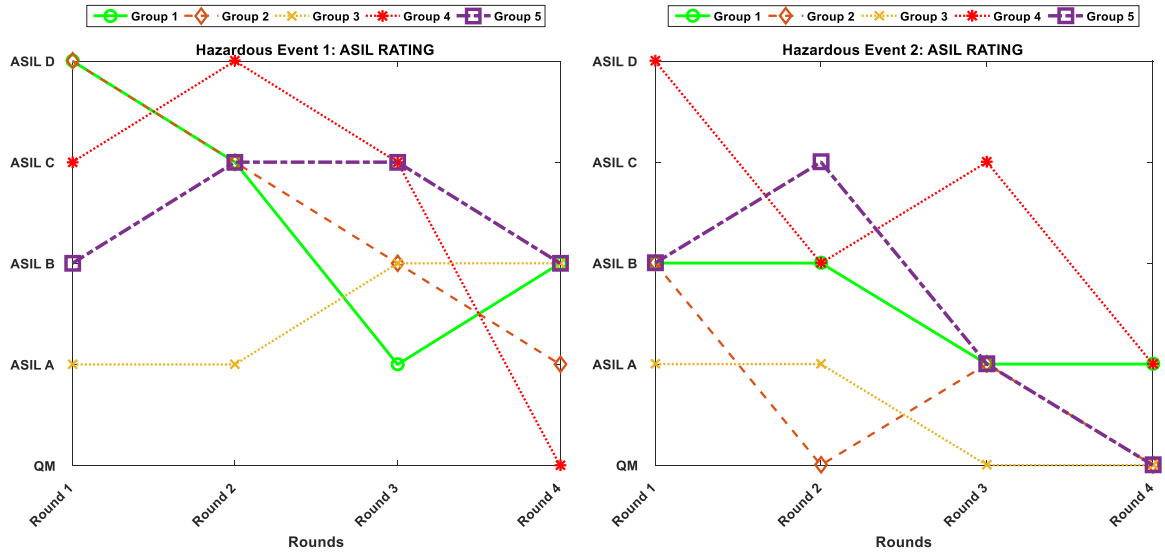

Figure A4.2: ASIL ratings in workshop 2 (Sweden)

Severity

In round 1, for both the hazardous events, the five groups gave three different ratings with the maximum difference between the ratings being in the order of two (S1 – S3) (Figure A4.3). Before round two, participants of the five groups were shuffled. In round two, for hazardous event 1, while four groups converged in their severity rating to S3, the fifth group differed by an order of one and gave S2 rating. As discussed earlier, in round three, three groups (groups 1, 2 and 3) were introduced to the rule-set for HARA and were asked to use the provided rule-set to perform HARA. The two other groups (group 4 and group 5) didn't receive any rules and the participants were shuffled among the groups.

In round 3, for hazardous event 1, out of the three groups which received the rule-set, two of the groups agreed on their severity rating of S3, while the third group's rating differed by an order of one (S2). The two other groups (group 4 and group 5) differed in their ratings by an order of one (S3 – S2). For hazardous event 2, the three groups using the rule-set converged in their severity rating at S2. However, the two groups (group 4 and group 5) who didn't use the rule-set differed in their rating by an order of two (S3 – S1).

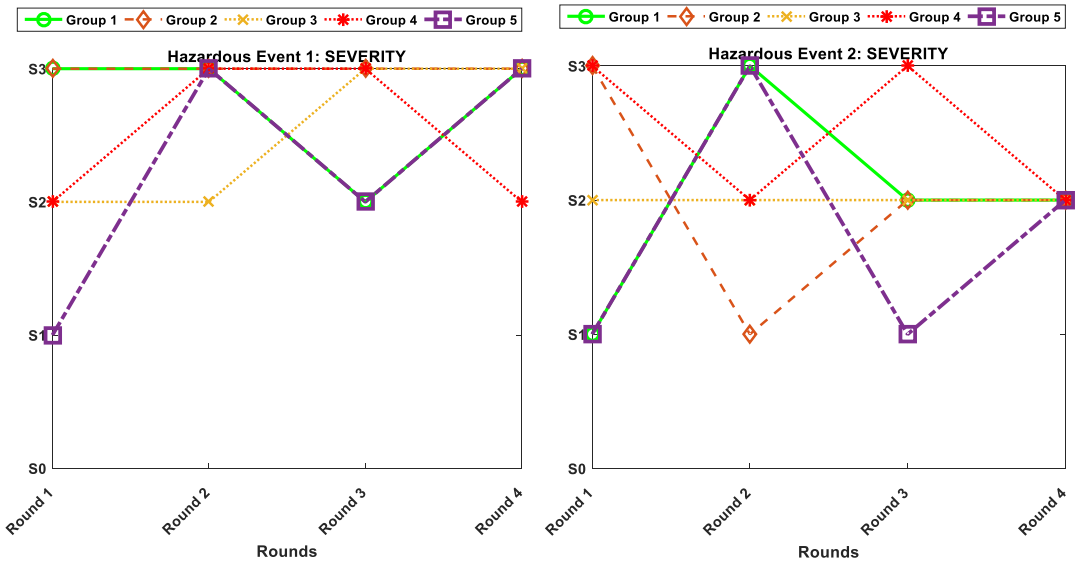

Figure A4.3: Severity ratings for workshop 2 (Sweden)

In round four, all participants from all five groups were shuffled and were told to use the rule-set provided to them for performing the HARA process. For hazardous event 1, four of the groups converged in their rating to S3, while the fifth group differed by an order of one (S2). Interestingly, for hazardous event 2, all groups converged in their severity rating to S2. Thus, suggesting that the rule-set potentially led to the convergence.

Exposure

Exposure ratings demonstrated an interesting trend for the two hazardous events. In round 1 and 2, for hazardous event 1, all five groups converged in their exposure ratings on E4 (Figure A4.4). However, for hazardous event 2, in round 1, while four of the groups converged on the exposure rating of E4, the fifth group differed by an order of one at E3. In round two, for hazardous event 2, the five groups gave three different ratings with a maximum difference in ratings of the order of two (E4 – E2).

In round three, for hazardous event 1 and hazardous event 2, the three groups which used the rule-set converged in their ratings for both the events, giving an E2 rating. However, the two other groups gave a rating of E4 for both the hazardous events. For round four, in which all five groups were told to use the rule-set, all five groups converged in their exposure rating and gave a rating of E2 for both the hazardous events.

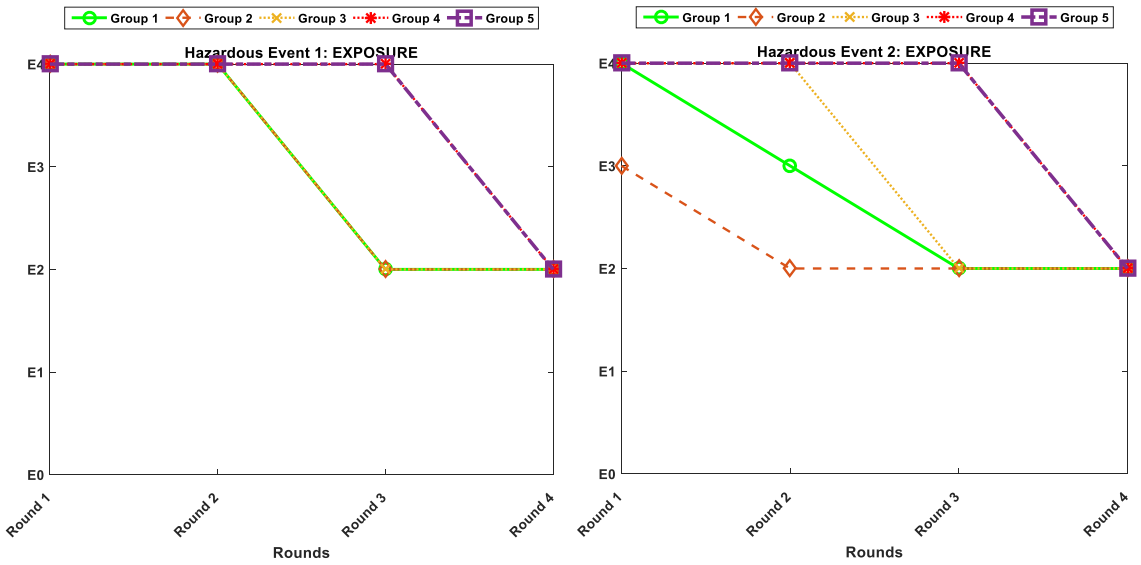

Figure A4.4: Exposure ratings for workshop 2 (Sweden)

Controllability

In workshop 1 (chapter eight), it was observed that among the ASIL components, controllability ratings tend to show the maximum inter-rater variation. A similar trend was observed in workshop 2 also (Figure A4.5). In round one, for hazardous event 1, four of the five groups gave a rating of C3 while the fifth group gave a rating of C1, differing by an order of two (C3 – C1). In round two, the five groups gave three different controllability ratings, with the maximum difference being of the order of two (C3 – C1).

For hazardous event 2, in round 1, the five groups gave three different ratings, with a maximum difference in ratings being of the order of two (C3 – C1). In round 2, while three groups converged in their rating on C3, two other groups differed by an order of one and gave a rating of C2.

In round three, for hazardous event 1, the three who were told to use the rule-set to assign their rating, converged to a controllability rating of C3. However, the two other groups, differed in their ratings by an order of one (C3 – C2). For hazardous event 2, the three groups using the rule-set gave two different ratings which differed from each other by an order of one (C3 – C2). The two other groups gave a rating of C2.

In round four, in which participants of all the groups were mixed and were told to use the rule-set, for hazardous event 1, four out of the five group converged in their rating at C3. However, the fifth group differed by an order of one (C3 – C2). Controllability rating for hazardous event 2 showed a large variation in round 4, even with the rule-set. The five

groups gave three different ratings with a maximum difference between the ratings being of the order of two (C3 – C1).

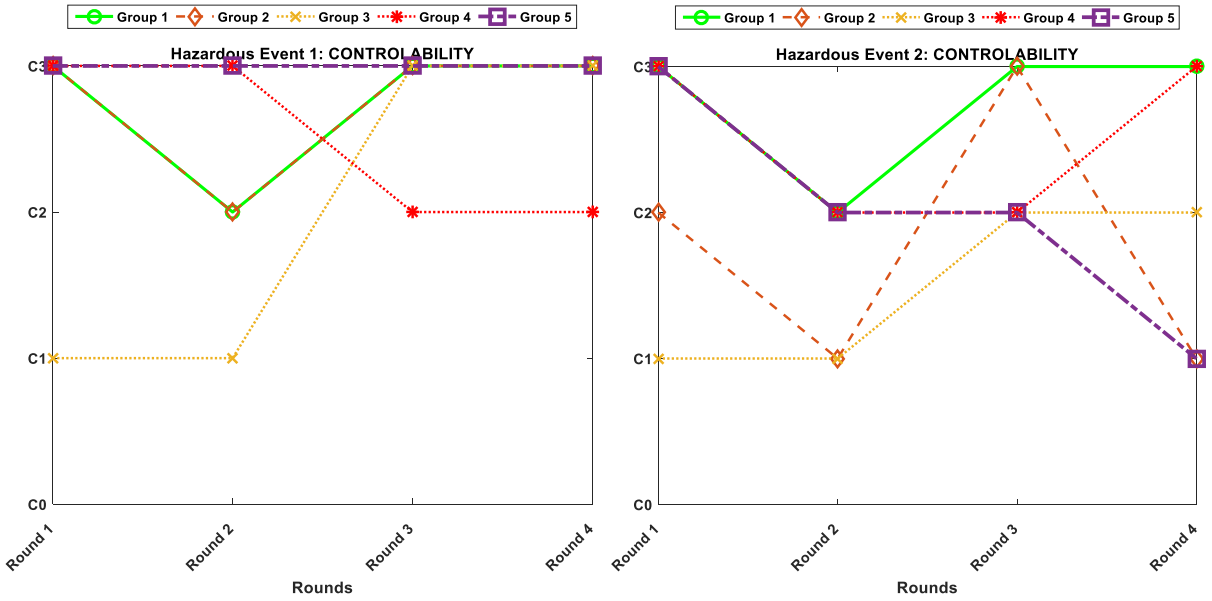

Figure A4.5: Controllability ratings for workshop 2 (Sweden)

A4.1.4.2. Qualitative results

Similar to workshop 1 (conducted in US, chapter eight), each of the five groups was asked to provide feedback to the following questions:

- (During the workshop) Have you experienced variation in hazard analysis discussions based on the group of people involved in the discussion?
- Do you think by having rules by parametrizing hazard analysis, we can have a more objective approach?

In response to the first question on whether the groups experienced variation in HARA, all five groups acknowledged high degree of variation. Furthermore, they commented that the variation was caused due to different assumptions made by the groups and also due to different level of experience among the members of the group. This is in line with the literature discussed in chapter six and the feedback received in workshop 1 (chapter eight). One of the participants commented: *"analysis… affected by the assumptions in regard to the operational situation. Could become very subjective without an underlying framework"*.

When asked about whether the introduction of rules improved the reliability of the automotive HARA process, all five groups responded positively and agreed that the rule-set

based on parametrisation of the ratings made the ratings more reliable. The groups suggested that the rule-set provided added guidance and helped support reasoning for the ratings. On the benefits of the parametrised approach, one of the participants commented: *"being able to compare hazard analysis between companies would be good"*.

However, all the groups mentioned that the rules needed to be more exhaustive to avoid the risk of omission of parameters in the analysis. One of the participants commented: *"you could accidently adapt your event description to tables and the rules available"*, while another participant said: *"easy to just look at the tales and stop thinking. Possible to miss hazards"*. More pointed feedback suggested the introduction of "type of impact" as an additional parameter for severity ratings. It was also suggested that the controllability rule-set suffered from over dimensionality as it had both TTC and vehicle velocity as parameters. Additionally, one of the participants mentioned that more clarity could be added by having separate tables for acceleration and deceleration.

The observed variation in ratings despite the introduction of the rule-set could be explained from the feedback from the qualitative groups. Most of the groups commented that the rule-set was not exhaustive and more parameters were needed. One of the groups commented: *"if parameters can be agreed upon, it would be beneficial"*, while another group commented: *"tables could be improved"* and *"rule-set would however need to be better specified"*. As the groups felt that some of the parameters were missing, the different groups made different assumptions for those parameters. Additionally, some of the groups found it hard to understand how to apply the rules (despite an initial example demo), with one of the groups suggesting that a *"user guide (at hand) could be necessary"*.

References

Abimbola, M., Khan, F., Khakzad, N., 2016. Risk-based safety analysis of well integrity operations. Saf. Sci. 84, 149–160. https://doi.org/10.1016/j.ssci.2015.12.009

Alexander, I., 2003. Misuse cases: Use cases with hostile intent. IEEE Softw. 20, 58–66. https://doi.org/10.1109/MS.2003.1159030

Altinger, H., Wotawa, F., Schurius, M., 2014. Testing Methods Used in the Automotive Industry : Results from a Survey, in: Proc. of the 2014 Workshop on Joining AcadeMiA and Industry Contributions to Test Automation and Model-Based Testing. ACM, 2014.

Ashleigh, M.J., Stanton, N.A., 2001. Trust : Key Elements in Human Supervisory Control Domains. Cogn. Technol. Work 3, 92–100.

Aven, T., 2015. Implications of black swans to the foundations and practice of risk assessment and management. Reliab. Eng. Syst. Saf. 134, 83–91. https://doi.org/10.1016/j.ress.2014.10.004

Aven, T., 2013. Safety Scien ce On the meaning of a black swan in a risk context. Saf. Sci. 57, 44–51. https://doi.org/10.1016/j.ssci.2013.01.016

Aven, T., 2010a. On how to define , understand and describe risk. Reliab. Eng. Syst. Saf. 95, 623–631. https://doi.org/10.1016/j.ress.2010.01.011

Aven, T., 2010b. On the Need for Restricting the Probabilistic Analysis in Risk Assessments to Variability 30, 354–360. https://doi.org/10.1111/j.1539-6924.2009.01314.x

Aven, T., Heide, B., 2009. Reliability and validity of risk analysis. Reliab. Eng. Syst. Saf. 94, 1862–1868. https://doi.org/10.1016/j.ress.2009.06.003

Aven, T., Reniers, G., 2013. How to define and interpret a probability in a risk and safety setting. Saf. Sci. 51, 223–231. https://doi.org/10.1016/j.ssci.2012.06.005

Aven, T., Zio, E., 2014. Foundational Issues in Risk Assessment and Risk Management 34, 1164–1172. https://doi.org/10.1111/risa.12132

Bainbridge, L., 1983. Ironies of automation. Automatica 19, 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Baker, S.P., O'Neill, B., Haddon, W., Long, W.B., 1974. The Injury Severity Score: A method for describing patients with multiple injuries and evaluating emergency care. J. Trauma 14.

Baldwin, C.L., Lewis, B.A., 2014. Perceived urgency mapping across modalities within a driving context. Appl. Ergon. 45, 1270–1277. https://doi.org/10.1016/j.apergo.2013.05.002

Balfe, N., Sharples, S., Wilson, J.R., 2015. Impact of automation : Measurement of performance , workload and behaviour in a complex control environment. Appl. Ergon. 47, 52–64. https://doi.org/10.1016/j.apergo.2014.08.002

Banks, V.A., Stanton, N.A., 2015. Keep the driver in control : Automating automobiles of the future. Appl. Ergon. 1–7. https://doi.org/10.1016/j.apergo.2015.06.020

Banks, V.A., Stanton, N.A., Harvey, C., 2014. Sub-systems on the road to vehicle automation : Hands and feet free but not ' mind ' free driving. Saf. Sci. 62, 505–514. https://doi.org/10.1016/j.ssci.2013.10.014

Bareket, Z., Fancher, P.S., Peng, H., Lee, K., Assaf, C.A., 2003. Methodology for Assessing Adaptive Cruise Control Behavior. IEEE Trans. Intell. Transp. Syst. 4, 123–131.

Bates, J., Leibling, D., 2012. Spaced Out Perspectives on parking policy.

Baysari, M.T., Caponecchia, C., Mcintosh, A.S., Wilson, J.R., 2009. Classification of errors contributing to rail incidents and accidents : A comparison of two human error identification techniques. Saf. Sci. 47, 948–

957. https://doi.org/10.1016/j.ssci.2008.09.012

BEA, 2012. Final Report: On the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris.

Beggiato, M., Krems, J.F., 2013. The evolution of mental model , trust and acceptance of adaptive cruise control in relation to initial information. Transp. Res. Part F Psychol. Behav. 18, 47–57. https://doi.org/10.1016/j.trf.2012.12.006

Beggiato, M., Pereira, M., Petzoldt, T., Krems, J., 2015. Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. Transp. Res. Part F Traffic Psychol. Behav. 35, 75–84. https://doi.org/10.1016/j.trf.2015.10.005

Beller, J., Heesen, M., Vollrath, M., 2013. Improving the Driver-Automation Interaction: An Approach Using Automation Uncertainty. Hum. Factors 55, 1130–1141. https://doi.org/10.1177/0018720813482327

Benmimoun, M., Pütz, A., Zlocki, A., Eckstein, L., 2012. euroFOT: Field Operational Test and Impact Assessment of Advanced Driver Assistance Systems: Final Results, in: Proceedings of the FISITA 2012 World Automotive Congress: Volume 9: Automotive Safety Technology. pp. 537–547.

Bennett, K.B., 2017. Ecological interface design and system safety : One facet of Rasmussen ' s legacy. Appl. Ergon. 59, 625–636. https://doi.org/10.1016/j.apergo.2015.08.001

Beukel, A.P. Van Den, Voort, M.C. Van Der, 2017. How to assess driver ' s interaction with partially automated driving systems e A framework for early concept assessment. Appl. Ergon. 59, 302–312. https://doi.org/10.1016/j.apergo.2016.09.005

BFU, 2004. Investigation Report of Ueberlingen Mid-Air Collision Accident.

Biassoni, F., Ruscio, D., Ciceri, R., 2016. Limitations and automation . The role of information about device-specific features in ADAS acceptability. Saf. Sci. 85, 179–186. https://doi.org/10.1016/j.ssci.2016.01.017

Billings, C.E., 1991a. Toward a Human-Centered Aircraft Automation Philosophy. Int. J. Aviat. Psychol. 1, 261–270. https://doi.org/10.1207/s15327108ijap0104

Billings, C.E., 1991b. Human-Centered Aircraft Automation: A Concept and Guidelines. California.

Biondi, F., Strayer, D.L., Rossi, R., Gastaldi, M., Mulatti, C., 2017. Advanced driver assistance systems : Using multimodal redundant warnings to enhance road safety. Appl. Ergon. 58, 238–244. https://doi.org/10.1016/j.apergo.2016.06.016

Bishop, R., 2000. A Survey of Intelligent Vehicle Applications Worldwide, in: Proc. of the IEEE Intelligent Vehicles Symposium 2000. Dearborn, Michigan, USA.

Björnsson, I., 2017. Holistic approach for treatment of accidental hazards during conceptual design of bridges - A case study in Sweden. Saf. Sci. 91, 168–180. https://doi.org/10.1016/j.ssci.2016.08.009

Blissing, B., Bruzelius, F., Ölvander, J., 2013. Augmented and Mixed Reality as a tool for evaluation of Vehicle Active Safety Systems, in: Proc. of the Road Safety and Simulation (RSS) International Conference 2013, Rome , Italy. Rome.

Bock, T., Maurer, M., 2007. Validation of the Vehicle in the Loop (VIL) – A milestone for the simulation of driver assistance systems 612–617.

Bock, T., Maurer, M., Färber, G., 2007. Validation of the Vehicle in the Loop (VIL) – A milestone for the simulation of driver assistance systems, in: Proc. of the 2007 IEEE Intelligent Vehicles Symposium (IV), Istanbul, Turkey. Istanbul.

Bock, T., Siedersberger, K.H., Maurer, M., 2005. Vehicle in the Loop - Augmented Reality Application for Collision Mitigation Systems, in: Proc. of the 4th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'05) 2005, Vienna, Austria.

Bonnefon, J., Shariff, A., Rahwan, I., 2016. The social dilemma of autonomous vehicles. Science (80-.). 352, 1573–1576. https://doi.org/10.1126/science.aaf2654

Cagno, E., Caron, F., Mancini, M., 1960. Multilevel Hazop for Risk Analysis in Plant Commissioning 77, 309–323.

Cairns, S., Harmer, C., Hopkin, J., Skippon, S., 2014. Sociological perspectives on travel and mobilities : A review. Transp. Res. Part A 63, 107–117. https://doi.org/10.1016/j.tra.2014.01.010

Cambridge English Dictionary [WWW Document], 2017. URL http://dictionary.cambridge.org/dictionary/english/ (accessed 3.3.17).

Campbell, R.L., 1992. Will the real scenario please stand up? ACM SIGCHI Bull. 24, 6–8. https://doi.org/10.1145/142386.1054872

Carbaugh, J., Godbole, D.N., Sengupta, R., 1998. Safety and capacity analysis of automated and manual highway systems 6, 69–99.

Carmines, E.G., Zeller, R.A., 1979. Reliability and Validity Assessment. Beverly Hills ; London : Sage Publications.

Casner, S.M., Hutchins, E.L., Norman, D., 2016. The Challenges of Partially Automated Driving. Commun. ACM 59, 70–77. https://doi.org/10.1145/2830565

Chapman, P., Underwood, G., 2000. Forgetting Near-Accidents : The Roles of Severity , Culpability and Experience in the Poor Recall of Dangerous Driving Situations. Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn. 14, 31–44.

Charette, R.N., 2009. This car runs on code. IEEE Spectr. 46.

Chavaillaz, A., Wastell, D., Sauer, J., 2016a. Effects of extended lay-off periods on performance and operator trust under adaptable automation. Appl. Ergon. 53, 241–251. https://doi.org/10.1016/j.apergo.2015.10.006

Chavaillaz, A., Wastell, D., Sauer, J., 2016b. System reliability , performance and trust in adaptable automation. Appl. Ergon. 52, 333–342. https://doi.org/10.1016/j.apergo.2015.07.012

Chen, S.T., Wall, A., Davies, P., Yang, Z., Wang, J., Chou, Y.H., 2013. A Human and Organisational Factors (HOFs) analysis method for marine casualties using HFACS-Maritime Accidents (HFACS-MA). Saf. Sci. 60, 105–114. https://doi.org/10.1016/j.ssci.2013.06.009

Christophe, J., Coze, L., 2013. What have we learned about learning from accidents ? Post-disasters reflections. Saf. Sci. 51, 441–453. https://doi.org/10.1016/j.ssci.2012.07.007

Cicchino, J.B., 2017. Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. Accid. Anal. Prev. 99, 142–152. https://doi.org/10.1016/j.aap.2016.11.009

Cockburn, A., 1997. Structuring use cases with goals. J. Object Oriented Program. 1997, 1–16.

Cockburn, A., Fowler, M., 1998. Question Time ! about Use Cases. ACM Sigplan Not. 33, 226–229.

Colonna, P., Intini, P., Berloco, N., Ranieri, V., 2016. The influence of memory on driving behavior : How route familiarity is related to speed choice . An on-road study. Saf. Sci. 82, 456–468. https://doi.org/10.1016/j.ssci.2015.10.012

Cullen, R.H., Rogers, W.A., Fisk, A.D., 2013. Human performance in a multiple-task environment : Effects of automation reliability on visual attention allocation. Appl. Ergon. 44, 962–968. https://doi.org/10.1016/j.apergo.2013.02.010

Dadashi, N., Stedmon, A.W., Pridmore, T.P., 2013. Semi-automated CCTV surveillance: The effects of system confidence, system accuracy and task complexity on operator vigilance, reliance and workload. Appl. Ergon. 44, 730–738. https://doi.org/10.1016/j.apergo.2012.04.012

Dambock, D., Weissgerber, T., Kienle, M., Bengler, K., 2013. Requirements for cooperative vehicle guidance, in: Proc. of the 16th IEEE Conference on Intelligent Transportation Systems (ITSC 2013). The Hague, The Netherlands, pp. 1656–1661. https://doi.org/10.1109/ITSC.2013.6728467

Damböck, D.D., Farid, M., Tönert, L., Bengler, K., Farid, D.M., Tönert, D.L., Beng-, P.K., 2012.

Übernahmezeiten beim hochautomatisierten Fahren. Tagung Fahrerassistenz 1–12.

Daziano, R.A., Sarrias, M., Leard, B., 2017. Are consumers willing to pay to let cars drive for them? Analyzing response to autonomous vehicles. Transp. Res. Part C Emerg. Technol. 78, 150–164. https://doi.org/10.1016/j.trc.2017.03.003

De Freitas, J., Anthony, S.E., Alvarez, G., 2019. Doubting Driverless Dilemmas.

Dekker, S., 2011. The criminalization of human error in aviation and healthcare : A review. Saf. Sci. 49, 121–127. https://doi.org/10.1016/j.ssci.2010.09.010

Department for Transport HM Government, 2015. Reported Road Casualties Great Britain: 2014 Annual Report. London, UK.

Department of Defence - US Govt., 1980. Procedures for Performing a Faliure Mode, Effects and Criticality Analysis: MIL-STD-1629A. Washington, DC.

DfT, 2017. The Highway Code [WWW Document].

Diels, C., Bos, J.E., 2015. Self-driving carsickness. Appl. Ergon. 53, 374–382. https://doi.org/10.1016/j.apergo.2015.09.009

Dogan, E., Rahal, M., Deborne, R., Delhomme, P., Kemeny, A., Perrin, J., 2017. Transition of control in a partially automated vehicle : Effects of anticipation and non-driving-related task involvement. Transp. Res. Part F Psychol. Behav. 46, 205–215. https://doi.org/10.1016/j.trf.2017.01.012

Duckworth, H.A., Moore, R.A., 2010. Social Responsibility : Failure Mode Effects and Analysis. CRC Press/Taylor & Francis, Boca Raton, FL.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P., 2003. The role of trust in automation reliance 58, 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Eichelberger, A.H., McCartt, A.T., 2014. Volvo drivers' experiences with advanced crash avoidance and related technologies. Traffic Inj. Prev. 15, 187–195. https://doi.org/10.1080/15389588.2013.798409

Ellims, M., Monkhouse, H.E., 2012. Agonising over ASILs: Controllability and the In-Wheel Motor, in: Proc. of the 7th IET International Conference on System Safety.

Endsley, M.R., Kiris, E.O., 1995. The Out-of-the-Loop Performance Problem and Level of Control in Automation 37, 381–394.

Ergai, A., Cohen, T., Sharp, J., Wiegmann, D., Gramopadhye, A., Shappell, S., 2016. Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and inter-rater reliability. Saf. Sci. 82, 393–398. https://doi.org/10.1016/j.ssci.2015.09.028

Ericson II, C.A., 2005. Hazard Analysis Techniques for System Safety. Wiley-Interscience, Hoboken, NJ.

Eriksson, A., Banks, V.A., Stanton, N.A., 2017. Transition to manual : Comparing simulator with on-road control transitions. Accid. Anal. Prev. 102, 227–234. https://doi.org/10.1016/j.aap.2017.03.011

Eriksson, A., Stanton, N.A., 2017a. Takeover Time in Highly Automated Vehicles. Hum. Factors 59, 689–705. https://doi.org/10.1177/0018720816685832

Eriksson, A., Stanton, N.A., 2017b. The chatty co-driver : A linguistics approach applying lessons learnt from aviation incidents. Saf. Sci. https://doi.org/10.1016/j.ssci.2017.05.005

Ervin, R.D., Sayer, J., LeBlanc, D., Bogard, S., Mefford, M., Hagan, M., Winkler, C., 2005. Automotive Collision Avoidance System Field Operational Test Report: Methodology and Results (DOT HS 809 900).

Fagnant, D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles : opportunities , barriers and policy recommendations. Transp. Res. Part A 77, 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Fagnant, D.J., Kockelman, K.M., 2014. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. Transp. Res. Part C Emerg. Technol. 40, 1–13. https://doi.org/10.1016/j.trc.2013.12.001

Faiella, G., Parand, A., Dean, B., Chana, P., Cesarelli, M., Stanton, N.A., Sevdalis, N., 2018. Expanding

healthcare failure mode and effect analysis : A composite proactive risk analysis approach. Reliab. Eng. Syst. Saf. 169, 117–126. https://doi.org/10.1016/j.ress.2017.08.003

Feldhütter, A., Gold, C., Hüger, A., Bengler, K., 2016. Trust in Automation as a matter of media and experience of automated vehicles. Proc. Hum. Factors Ergon. Soc. 60th Annu. Meet. 2024–2028.

Fitts, P.M., Chapanis, A., Frick, F.C., Garner, W.R., Gebhard, J.W., Grether, W.F., Henneman, R.H., Kappauf, W.E., Newman, E.B., A.C.Williams, J., 1951. Human engineering for an effective air - navigation and traffic - control system. Washington, D.C., USA.

Flage, R., Aven, T., 2015. Emerging risk – Conceptual de fi nition and a relation to black swan type of events. Reliab. Eng. Syst. Saf. 144, 61–67. https://doi.org/10.1016/j.ress.2015.07.008

Fleming, C.H., Spencer, M., Thomas, J., Leveson, N., Wilkinson, C., 2013. Safety assurance in NextGen and complex transportation systems. Saf. Sci. 55, 173–187. https://doi.org/10.1016/j.ssci.2012.12.005

Fouche, C., Light, G., 2011. An Invitation to Dialogue: "The World Cafe" In Social Work Research. Qual. Soc. Work 10, 28–48. https://doi.org/10.1177/1473325010376016

France, M.E., 2017. Engineering for Humans : A New Extension to STPA. MIT.

Fuller, R., 2005. Towards a general theory of driver behaviour. Accid. Anal. Prev. 37, 461–472. https://doi.org/10.1016/j.aap.2004.11.003

Gietelink, O.J., Labibes, K., Verburg, D.J., Oostendorp, A.F., 2004. Pre-crash system validation with PRESCAN and VEHIL, in: Proc. of the IEEE Intelligent Vehicles Symposium (IV), 2004, Parma, Italy. Parma, Italy, pp. 913–918. https://doi.org/10.1109/IVS.2004.1336507

Goerlandt, F., Khakzad, N., Reniers, G., 2016. Validity and validation of safety-related quantitative risk analysis : A review. Saf. Sci. https://doi.org/10.1016/j.ssci.2016.08.023

Goerlandt, F., Montewka, J., 2015. A framework for risk analysis of maritime transportation systems : A case study for oil spill from tankers in a ship – ship collision. Saf. Sci. 76, 42–66. https://doi.org/10.1016/j.ssci.2015.02.009

Goerlandt, F., Reniers, G., 2016. On the assessment of uncertainty in risk diagrams. Saf. Sci. 84, 67–77. https://doi.org/10.1016/j.ssci.2015.12.001

Gold, C., Damböck, D., Lorenz, L., Bengler, K., 2013. "Take over!" How long does it take to get the driver back into the loop? Proc. Hum. Factors Ergon. Soc. Annu. Meet. 57, 1938–1942. https://doi.org/10.1177/1541931213571433

Gordon, T., Sardar, H., Blower, D., Aust, L.M., Bareket, Z., Barnes, M., Blankespoor, A., Isaksson-Hellman, I., Ivarsson, J., Juhas, B., Nobukawa, K., Theander, H., 2010. Advanced Crash Avoidance Technologies (ACAT) Program – Final Report of the Volvo-Ford- UMTRI Project: Safety Impact Methodology for Lane Departure Warning – Method Development And Estimation of Benefits (DOT HS 811 405).

Green, M., 2000. " How Long Does It Take to Stop ?" Methodological Analysis of Driver Perception-Brake Times 2, 195–216.

Guériau, M., Billot, R., El Faouzi, N.E., Monteil, J., Armetta, F., Hassas, S., 2016. How to assess the benefits of connected vehicles? A simulation framework for the design of cooperative traffic management strategies. Transp. Res. Part C Emerg. Technol. 67, 266–279. https://doi.org/10.1016/j.trc.2016.01.020

Haboucha, C.J., Ishaq, R., Shiftan, Y., 2017. User preferences regarding autonomous vehicles. Transp. Res. Part C 78, 37–49. https://doi.org/10.1016/j.trc.2017.01.010

Hancock, P.A., 2014. Automation: how much is too much? Ergonomics 57, 449–454. https://doi.org/10.1080/00140139.2013.816375

Hansson, S.O., Aven, T., 2014. Is Risk Analysis Scientific ? 34. https://doi.org/10.1111/risa.12230

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, in: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload.

https://doi.org/10.1016/S0166-4115(08)62386-9

Haslbeck, A., Zhang, B., 2017. I spy with my little eye : Analysis of airline pilots ' gaze patterns in a manual instrument fl ight scenario. Appl. Ergon. 63, 62–71. https://doi.org/10.1016/j.apergo.2017.03.015

Helldin, T., Falkman, G., Riveiro, M., Davidsson, S., 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. Proc. Int. Conf. Automot. User Interfaces Interact. Veh. Appl. - AutomotiveUI '13 5, 210–217. https://doi.org/10.1145/2516540.2516554

Hendriks, F., Pelders, R., Tideman, M., 2010. Future Testing of Active Safety Systems, 3rd ed, SAE International Journal of Passenger Cars - Electronic and Electrical Systems. SAGE. https://doi.org/10.4271/2010-01-2334

Hergeth, S., Lorenz, L., Krems, J.F., 2017. Prior Familiarization With Takeover Requests Affects Drivers' Takeover Performance and Automation Trust. Hum. Factors 59, 457–470. https://doi.org/10.1177/0018720816678714

Hill, S.G., Lavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklad, A.L., Christ, R.E., 1992. Comparison of four subjective workload rating scales. Hum. Factors 34, 429–439. https://doi.org/10.1177/001872089203400405

Hoedemaeker, M., 2000. Driving with intelligent vehicles: Driving behaviour with ACC and the acceptance by individual drivers, in: Proc. of the IEEE Intelligent Transportation Systems Conference (ITSC2000). Dearborn, Michigan, USA, pp. 506–509. https://doi.org/10.1109/ITSC.2000.881121

Hoffman, R.R., Lintern, G., Eitelman, S., 2004. The Janus Principle. IEEE Intell. Syst. 19, 78–80. https://doi.org/10.1109/MIS.2004.1274915

Huang, W., Wang, K., Lv, Y., Zhu, F., 2016. Autonomous Vehicles Testing Methods Review, in: Proc. of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC) 2016. IEEE, Rio de Janiero, pp. 163–168. https://doi.org/10.1109/ITSC.2016.7795548

IEC, 2016. Hazard and operability studies (HAZOP studies) — Application guide: IEC 61882.

IEEE, 2010. Systems and software engineering - Vocabulary.(ISO/IEC/IEEE 24765).

Inagaki, T., 2003. Adaptive Automation : Sharing and Trading of 147–169.

Ishimatsu, T., Leveson, N., Thomas, J., Katahira, M., Miyamoto, Y., Nakao, H., 2010. Modeling and Hazard Analysis Using Stpa, in: Proc. of the 4th IAASS Conference, Making Safety Matter, 19–21 May 2010. Huntsville, Alabama, USA.

ISO, 2018a. Road vehicles — Functional safety - Part 11 (ISO 26262).

ISO, 2018b. Road vehicles — Functional safety (ISO 26262).

ISO, 2018c. Road vehicles — Functional safety - Part 3 (ISO 26262).

ISO, 2018d. Road vehicles — Functional safety - Part 6 (ISO 26262).

ISO, 2013. Software and systems engineering — Software testing - Part 1: Concepts and definitions (ISO/IEC/IEEE 29119-1).

ISO, 2011a. Road vehicles — Functional safety (ISO 26262). SAGE.

ISO, 2011b. Road vehicles — Functional safety (ISO 26262) Part 3 : Concept phase. SAGE.

Itoh, M., Abe, G., Tanaka, K., 1999. Trust in and Use of Automation: Their Dependence on Occurrence Patterns of Malfunctions, in: Proc. of the 1999 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, Tokyo, Japan, pp. 715–720.

Itoh, M., Horikome, T., Inagaki, T., 2013. Effectiveness and driver acceptance of a semi-autonomous forward obstacle collision avoidance system. Appl. Ergon. 44, 756–763. https://doi.org/10.1016/j.apergo.2013.01.006

Jakus, G., Dicke, C., Sodnik, J., 2015. A user study of auditory , head-up and multi-modal displays in vehicles. Appl. Ergon. 46, 184–192. https://doi.org/10.1016/j.apergo.2014.08.008

Jamson, A.H., Merat, N., Carsten, O.M.J., Lai, F.C.H., 2013. Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions. Transp. Res. Part C Emerg. Technol. 30, 116–125. https://doi.org/10.1016/j.trc.2013.02.008

Jenness, J.W., Lerner, N.D., Mazor, S.D., Osberg, J.S., Tefft, B.C., 2007. Use of Advanced In-Vehicle Technology by Young and Older Early Adopters. Survey Results on Navigation Systems.

Jeon, M., Walker, B.N., Gable, T.M., 2015. The effects of social interactions with in-vehicle agents on a driver's anger level, driving performance, situation awareness, and perceived workload. Appl. Ergon. 50, 185–199. https://doi.org/10.1016/j.apergo.2015.03.015

Jian, J.-Y., Bisantz, A.M., Drury, C.G., 2000. Foundations for an Empirically Determined Scale of Trust in Automated System. Int. J. Cogn. Ergon. 4, 53–71.

Johansson, R., 2009. Vision Zero - Implementing a Policy for Traffic Safety. Saf. Sci. 47, 826–831. https://doi.org/10.1016/j.ssci.2008.10.023

Johansson, R., Nilsson, J., 2016a. The Need for an Environment Perception Block to Address all ASIL Levels Simultaneously, in: Proc. of the IEEE Intelligent Vehicles Symposium 2016. Gothenburg.

Johansson, R., Nilsson, J., 2016b. Disarming the Trolley Problem – Why Self-driving Cars do not Need to Choose Whom to Kill, in: Proc. of the Workshop Critical Automotive Applications : Robustness & Safety. Gothenburg, Sweden.

Kaiser, B., Liggesmeyer, P., Mäckel, O., 2003. A New Component Concept for Fault Trees, in: Proc. of the 8th Australian Workshop on Safety Critical Systems and Software. Canberra, Australia, pp. 37–46.

Kalra, N., Groves, D.G., 2017. The Enemy of Good.

Kalra, N., Paddock, S.M., 2016a. Driving to safety : How many miles of driving would it take to demonstrate autonomous vehicle reliability ? Transp. Res. Part A 94, 182–193. https://doi.org/10.1016/j.tra.2016.09.010

Kalra, N., Paddock, S.M., 2016b. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transp. Res. Part A Policy Pract. 94, 182–193. https://doi.org/10.1016/j.tra.2016.09.010

Karltun, A., Karltun, J., Berglund, M., Eklund, J., 2017. HTO – A complementary ergonomics approach. Appl. Ergon. 59, 182–190. https://doi.org/10.1016/j.apergo.2016.08.024

Kazi, T.A., Stanton, N.A., Walker, G.H., Young, M.S., 2007. Designer driving: drivers' conceptual models and level of trust in adaptive cruise control. Int. J. Veh. Des. 45, 339–360.

Keeling, G., 2019. Why Trolley Problems Matter for the Ethics of Automated Vehicles. Sci. Eng. Ethics. https://doi.org/10.1007/s11948-019-00096-1

Kelly, T.P., 2004. A Systematic Approach to Safety Case Management, in: SAE Technical Paper: 2004-01-1779. pp. 257–266. https://doi.org/10.4271/2004-01-1779

Kelm, G.G., 2010. Failure Modes and Effects Analysis (FMEA), Critical Items List (CIL), and Fault Tree Analysis (FTA) - Work Instruction.

Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G., 1993. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. Int. J. Aviat. Psychol. 3, 203–220.

Khakzad, N., Khan, F., Amyotte, P., 2012. Dynamic risk analysis using bow-tie approach. Reliab. Eng. Syst. Saf. 104, 36–44. https://doi.org/10.1016/j.ress.2012.04.003

Khakzad, N., Khan, F., Paltrinieri, N., 2014. On the application of near accident data to risk analysis of major accidents. Reliab. Eng. Syst. Saf. 126, 116–125. https://doi.org/10.1016/j.ress.2014.01.015

Khastgir, S., Birrell, S., Dhadyalla, G., Jennings, P., 2017. Calibrating Trust to Increase the Use of Automated Systems in a Vehicle, in: Stanton, N., Landry, S., Bucchianico, G. Di, Vallicelli, A. (Eds.), Advances in Human Aspects of Transportation. Advances in Intelligent Systems and Computing. Springer, Cham, pp.

535–546.

Koopman, P., Wagner, M., 2016. Challenges in Autonomous Vehicle Testing and Validation. SAE Int. J. Transp. Saf. 4, 2016-01–0128. https://doi.org/10.4271/2016-01-0128

Körber, M., Gold, C., Lechner, D., Bengler, K., 2016. The influence of age on the take-over of vehicle control in highly automated driving 39, 19–32. https://doi.org/10.1016/j.trf.2016.03.002

Kyriakidis, M., Happee, R., De Winter, J.C.F., 2015. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. Transp. Res. Part F Traffic Psychol. Behav. 32, 127–140. https://doi.org/10.1016/j.trf.2015.04.014

Labib, A., 2015. Learning (and unlearning) from failures : 30 years on from Bhopal to Fukushima an analysis through reliability engineering techniques. Process Saf. Environ. Prot. 97, 80–90. https://doi.org/10.1016/j.psep.2015.03.008

Larsson, A.F.L., 2012. Driver usage and understanding of adaptive cruise control. Appl. Ergon. 43, 501–506. https://doi.org/10.1016/j.apergo.2011.08.005

Larsson, A.F.L., Kircher, K., Hultgren, J.A., 2014. Learning from experience: Familiarity with ACC and responding to a cut-in situation in automated driving. Transp. Res. Part F Traffic Psychol. Behav. 27, 229–237. https://doi.org/10.1016/j.trf.2014.05.008

Le Coze, J.C., 2013. New models for new times. An anti-dualist move. Saf. Sci. 59, 200–218. https://doi.org/10.1016/j.ssci.2013.05.010

Le, S., Liu, X., Zheng, F., Polak, J., 2016. Automated cars : Queue discharge at signalized intersections with ‘ Assured-Clear-Distance-Ahead ’ driving strategies. Transp. Res. Part C Emerg. Technol. 62, 35–54. https://doi.org/10.1016/j.trc.2015.11.005

Le, S., Zolfaghari, A., Polak, J., 2015. Autonomous cars : The tension between occupant experience and intersection capacity. Transp. Res. Part C Emerg. Technol. 52, 1–14. https://doi.org/10.1016/j.trc.2015.01.002

LeBlanc, D., Sayer, J., Winkler, C., Ervin, R., Bogard, S., Mefford, D.J., Hagan, M., Bareket, M., Z., Goodsell, R., Gordon, T., 2006. Road Departure Crash Warning System Field Operational Test: Methodology and Results.

Lee, J., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35, 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, J.D., See, K.A., 2004. Trust in Automation: Designing for Appropriate Reliance. Hum. Factors 46, 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lee, W.S., Grosh, D.L., Tillman, F.A., Lie, C.H., 1985. Fault Tree Analysis, Methods, and Applications - A Review. IEEE Trans. Reliab. R-34, 194–203. https://doi.org/10.1109/TR.1985.5222114

Lenné, M.G., Liu, C.C., Salmon, P.M., Holden, M., Moss, S., 2011. Minimising risks and distractions for young drivers and their passengers: An evaluation of a novel driver-passenger training program. Transp. Res. Part F Traffic Psychol. Behav. 14, 447–455. https://doi.org/10.1016/j.trf.2011.08.001

Lenné, M.G., Salmon, P.M., Liu, C.C., Trotter, M., 2012. A systems approach to accident causation in mining : An application of the HFACS method. Accid. Anal. Prev. 48, 111–117. https://doi.org/10.1016/j.aap.2011.05.026

Leveson, N., 2004. A new accident model for engineering safer systems 42, 237–270. https://doi.org/10.1016/S0925-7535(03)00047-X

Leveson, N.G., 2017. Rasmussen’s legacy: A paradigm change in engineering for safety. Appl. Ergon. 59, 581–591. https://doi.org/10.1016/j.apergo.2016.01.015

Leveson, N.G., 2012. Engineering a Safer World: Systems Thinking Applied to Safety, The MIT Press. https://doi.org/10.1136/injuryprev-2015-041920

Leveson, Nancy G., 2011. Engineering a Safer World. The MIT Press.

Leveson, Nancy G, 2011. Applying systems thinking to analyze and learn from events. Saf. Sci. 49, 55–64. https://doi.org/10.1016/j.ssci.2009.12.021

Leveson, N.G., 2006. New Safety Technologies for the Automotive Industry.

Leveson, N.G., Thomas, J.P., 2018. STPA Handbook. https://doi.org/10.2143/JECS.64.3.2961411

Lewandowsky, S., Mundy, M., Tan, G.P.A., 2000. The Dynamics of Trust: Comparing Humans to Automation. J. Exp. Pschology Appl. 6, 104–123.

Lions, J.L., 1996. Ariane 5 Flight 501 Failure: Report by the Inquiry Board.

Lorenz, L., Kerschbaum, P., Schumann, J., 2014. Designing take over scenarios for automated driving: How does augmented reality support the driver to get back into the loop? Proc. Hum. Factors Ergon. Soc. 58th Annu. Meet. - 2014 58, 1681–1685. https://doi.org/10.1177/1541931214581351

Lortie, M., Rizzo, P., 1998. The classification of accident data. Saf. Sci. 31, 31–57. https://doi.org/10.1016/S0925-7535(98)00053-8

Louise Barriball, K., While, A., 1994. Collecting data using a semi-structured interview: a discussion paper. J. Adv. Nurs. 19, 328–335. https://doi.org/10.1111/j.1365-2648.1994.tb01088.x

Louw, T., Madigan, R., Carsten, O., Merat, N., 2016. Were they in the loop during automated driving ? Links between visual attention and crash potential 1–6. https://doi.org/10.1136/injuryprev-2016-042155

Louw, T., Merat, N., 2017. Are you in the loop ? Using gaze dispersion to understand driver visual attention during vehicle automation. Transp. Res. Part C 76, 35–50. https://doi.org/10.1016/j.trc.2017.01.001

Lu, Z., Coster, X., Winter, J. De, 2017. How much time do drivers need to obtain situation awareness ? A laboratory-based study of automated driving. Appl. Ergon. 60, 293–304. https://doi.org/10.1016/j.apergo.2016.12.003

Ma, J., Zhou, F., Huang, Z., James, R., 2018. Hardware-in-the-Loop Testing of Connected and Automated Vehicle Applications : A Use Case for Cooperative Adaptive Cruise Control, in: 2Proc. of the 21st International Conference on Intelligent Transportation Systems (ITSC) 2018. IEEE, pp. 2878–2883.

Malek, S., 2017. What Software Developers Can Learn From the Latest Car Recalls [WWW Document]. CISQ.

Markey, the staff of S.E.J., 2015. Tracking & Hacking : Security & Privacy Gaps Put American Drivers at Risk. Senat. Edward J. Markey (D-Massaschussetts). 14.

Marsden, G., Mcdonald, M., Brackstone, M., 2001. Towards an understanding of adaptive cruise control 9, 33–51.

Martens, Marieke H, Fox, M., 2007. Does road familiarity change eye fixations ? A comparison between watching a video and real driving 10, 33–47. https://doi.org/10.1016/j.trf.2006.03.002

Martens, Marieke H., Fox, M.R.J., 2007. Do familiarity and expectations change perception? Drivers' glances and response to changes. Transp. Res. Part F Traffic Psychol. Behav. 10, 476–492. https://doi.org/10.1016/j.trf.2007.05.003

Merat, N., Jamson, a. H., Lai, F.C.H., Carsten, O., 2012. Highly Automated Driving, Secondary Task Performance, and Driver State. Hum. Factors J. Hum. Factors Ergon. Soc. 54, 762–771. https://doi.org/10.1177/0018720812442087

Michon, J.A., 1985. A Critical View of Driver Behaviour Models: What do we know, What should we do?, in: Schwing, L.E.& R.C. (Ed.), Human Behavior and Traffic Safety. Plenum Press, New York, pp. 485–520.

Miller, C., Valasek, C., 2015. Remote Exploitation of an Unaltered Passenger Vehicle 2015, 1–91.

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., Ju, W., 2016. Behavioral Measurement of Trust in Automation : The Trust Fall 1849–1853.

Mohebbi, R., Gray, R., Tan, H.Z., 2009. Driver Reaction Time to Tactile and Auditory Rear-End Collision Warnings While Talking on a Cell Phone. Hum. Factors J. Hum. Factors Ergon. Soc. 51, 102–110.

https://doi.org/10.1177/0018720809333517.

Molesworth, B.R.C., Koo, T.T.R., 2016. The influence of attitude towards individuals??? choice for a remotely piloted commercial flight: A latent class logit approach. Transp. Res. Part C Emerg. Technol. 71, 51–62. https://doi.org/10.1016/j.trc.2016.06.017

Monkhouse, H., Habli, I., Mcdermid, J., 2015. The Notion of Controllability in an autonmous vehicle context, in: CARS 2015 - Critical Automotive Applications: Robustness & Safety. Sep 2015. Paris, France.

Mosleh, A., Bier, V.M., Apostolakis, G., 1988. A Critique of Current Practice for the Use of Expert Opinions in Probabilistic Risk Assessment 20, 63–85.

Mouloua, M., Gilson, R., Kring, J., Hancock, P., 2001. Workload, Situation Awareness, and Teaming Issues for UAV/UCAV Operations, 3rd ed, Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE. https://doi.org/10.1177/154193120104500235

Muir, B.M., 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37, 1905–1922. https://doi.org/10.1080/00140139408964957

Muir, B.M., 1987. Trust between humans and machines , and the design of decision aids 527–539.

Muir, B.M., Moray, N., 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics 39, 429–460. https://doi.org/10.1080/00140139608964474

Nakajima, T.T., 2008. Advanced Driver Assist System: Its Challenges and Solutions from the Customer's Point of View, in: SAE Convergence 2008, Detroit, Michigan. SAE Paper No. 2008-21-0031. SAGE.

Naujoks, F., Mai, C., Neukum, A., 2014. The effect of urgency of take-over requests during highly automated driving under distraction conditions.

Naujoks, F., Purucker, C., Neukum, A., Wolter, S., Steiger, R., 2015. Controllability of Partially Automated Driving functions - Does it matter whether drivers are allowed to take their hands off the steering wheel? Transp. Res. Part F Traffic Psychol. Behav. 35, 185–198. https://doi.org/10.1016/j.trf.2015.10.022

Navarro, J., Deniel, J., Yousfi, E., Jallais, C., Bueno, M., Fort, A., 2017. Influence of lane departure warnings onset and reliability on car drivers' behaviors. Appl. Ergon. 59, 123–131. https://doi.org/10.1016/j.apergo.2016.08.010

NHTSA, 2017a. Investigation Report: PE 16-007 (MY2014-2016 Tesla Model S and Model X).

NHTSA, 2017b. PE 16-007.

Nicola, G., Pariota, L., Brackstione, M., Mcdonald, M., 2013. Driving behaviour models enabling the simulation of Advanced Driving Assistance Systems : revisiting the Action Point paradigm. Transp. Res. Part C 36, 352–366. https://doi.org/10.1016/j.trc.2013.09.009

Norman, D.A., 1990. The problem with automation: Inappropriate feedback and interaction, not over - automation. Philos. Trans. R. Soc. London, B 327, 585–593. https://doi.org/10.1098/rstb.1990.0101

NTSB, 2018. Preliminary Report HWY18MH010.

NTSB, 1988. Aircraft Accident Report - Northwest Airlines, Inc. McDonnell Douglas DC-9-82, N312RC, Detroit Metropolitian Wayne Country Airport, Romulus, Michigan, August 16, 1987 (NTSB/AAR-88/05). Washington, D.C., USA.

NTSB, 1986. China Airlines Boeing 747-SP, N4522V, 300 Nautical Miles Northwest of San Francisco, California, February 19, 1985. Report No. NTSB/AAR-86/03.

Olsen, N., Williamson, A., 2017. Application of classification principles to improve the reliability of incident classification systems: A test case using HFACS-ADF. Appl. Ergon. 63, 31–40. https://doi.org/10.1016/j.apergo.2017.03.014

Parasuraman, R., 1987. Human-Computer Monitoring. Hum. Factors 29, 695–706.

Parasuraman, R., Miller, C. a., 2004. Trust and etiquette in high-criticality automated systems. Commun. ACM 47, 51. https://doi.org/10.1145/975817.975844

Parasuraman, R., Molloy, R., Singh, I.L., 1993. Performance Consequences of Automation - Induced "Complacency." Int. J. Aviat. Psychol. 3, 1–23. https://doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., Riley, V., 1997. Humans and Automation : Use , Misuse , Disuse , Abuse 39, 230–253.

Pawlicki, T., Samost, A., Brown, D.W., Manger, R.P., Kim, G.Y., Leveson, N.G., 2016. Application of systems and control theory-based hazard analysis to radiation oncology. Med. Phys. 43, 1514–1530. https://doi.org/10.1118/1.4942384

Paxion, J., Galy, E., Berthelon, C., 2015. Overload depending on driving experience and situation complexity : Which strategies faced with a pedestrian crossing ? Appl. Ergon. 51, 343–349. https://doi.org/10.1016/j.apergo.2015.06.014

Peng, Y., Ng, L., Lee, J.D., 2014. Reading , typing , and driving : How interactions with in-vehicle systems degrade driving performance. Transp. Res. Part F Psychol. Behav. 27, 182–191. https://doi.org/10.1016/j.trf.2014.06.001

Pereira, M., Beggiato, M., Petzoldt, T., 2015. Use of adaptive cruise control functions on motorways and urban roads : Changes over time in an on-road study. Appl. Ergon. 50, 105–112. https://doi.org/10.1016/j.apergo.2015.03.002

Perrow, C., 2011. Normal Accidents: Living with High Risk Technologies. Princeton university press.

Perrow, C., 1981. Normal Accident at Three Mile Island.

Petermeijer, S., Bazilinskyy, P., Bengler, K., Winter, J. De, 2017. Take-over again : Investigating multimodal and directional TORs to get the driver back into the loop. Appl. Ergon. 62, 204–215. https://doi.org/10.1016/j.apergo.2017.02.023

Presidential Commision, 1986. Report to the President By the Presidential Commision On the Space Shuttle Challenger Accident.

Radlmayr, J., Gold, C., Lorenz, L., Farid, M., Bengler, K., 2014. How Traffic Situations and Non-Driving Related Tasks Affect the Take-Over Quality in Highly Automated Driving. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 58, 2063–2067. https://doi.org/10.1177/1541931214581434

Rae, A., Alexander, R., 2017. Forecasts or fortune-telling : When are expert judgements of safety risk valid ? Saf. Sci. 99, 156–165. https://doi.org/10.1016/j.ssci.2017.02.018

Rajaonah, B., Anceaux, F., Vienne, F., 2006. Trust and the use of adaptive cruise control: a study of a cut-in situation. Cogn. Technol. Work 8, 146–155. https://doi.org/10.1007/s10111-006-0030-3

Rajaonah, B., Tricot, N., Anceaux, F., Millot, P., 2008. The role of intervening variables in driver-ACC cooperation. Int. J. Hum. Comput. Stud. 66, 185–197. https://doi.org/10.1016/j.ijhcs.2007.09.002

Randolph, J., Simpson, A.K., Culver, J.C., Baker, H.H., Moynihan, D.P., Domenici, P. V, 1980. Nuclear Accident and Recovery at Three Mile Island.

Rasmussen, J., 1990. Human Error and the Problem of Causality in Analysis of Accidents. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 327, 449–462. https://doi.org/10.1098/rstb.1990.0088

Rasmussen, J., 1985. The Role of Hierarchical Knowledge Representation in Decisionmaking and System Management 234–243.

Rasmussen, J., 1983. Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models. IEEE Trans. Syst. Man. Cybern. 13, 257–266.

Rasmussen, J., 1982. Human errors. A taxonomy for describing human malfunction in industrial installations. J. Occup. Accid. 4, 311–333. https://doi.org/10.1016/0376-6349(82)90041-4

Reason, J., 1990. Human Error. Cambridge University Press.

Reay, K.A., Andrews, J.D., 2002. A fault tree analysis strategy using binary decision diagrams 78, 45–56.

Reimer, B., Mehler, B., Coughlin, J.F., 2016. Reductions in self-reported stress and anticipatory heart rate with the use of a semi-automated parallel parking system. Appl. Ergon. 52, 120–127.

https://doi.org/10.1016/j.apergo.2015.07.008

Rene, M., Klemm, S., Kuhnt, F., Schamm, T., Zollner, J.M., 2016. Testing and Validating High Level Components for Automated Driving : Simulation Framework for Traffic Scenarios, in: Proc. of the 2016 IEEE Intelligent Vehicles Symposium (IV). Gothenburg. https://doi.org/10.1109/IVS.2016.7535378

Richards, D., Stedmon, A., 2016. To delegate or not to delegate : A review of control frameworks for autonomous cars. Appl. Ergon. 53, 383–388. https://doi.org/10.1016/j.apergo.2015.10.011

Riley, D., 2014. Mental models in warnings message design : A review and two case studies. Saf. Sci. 61, 11–20. https://doi.org/10.1016/j.ssci.2013.07.009

Riley, V., 1996. Operator Reliance on Automation: Theory and Data, in: Parasuraman, R., Mouloua, M. (Eds.), Automation and Human Performance: Theory and Applications. Lawrence Erlbaum Associates, Publishers, pp. 19–35.

Robinson-mallett, C., Grochtmann, M., Wegener, J., Köhnlein, J., Kühn, S., 2010. Modelling requirements to support testing of product lines 11–18. https://doi.org/10.1109/ICSTW.2010.65

Robinson-mallett, C.L., 2012. An Approach on Integrating Models and Textual Specifications 92–96.

Robinson, G.H., 1986. Towards a Methodology for the Design of Warnings, in: Proc. of the Human Factors Society - 30th Annual Meeting 1986.

Robson, C., McCartan, K., 2016. Real world research : a resource for users of social research methods in applied settings, 4th ed. Wiley.

Rosqvist, T., 2010. On the validation of risk analysis — A commentary. Reliab. Eng. Syst. Saf. 95, 1261–1265. https://doi.org/10.1016/j.ress.2010.06.002

SAE International, 2018. Surface Vehicle Recommended Practice: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles - J3016. https://doi.org/10.4271/2012-01-0107.

SAE International, 2016a. SAE J3114 - Human Factors Definitions for Automated Driving and Related Research Topics.

SAE International, 2016b. Surface Vehicle Recommended Practice: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.

SAE International, 2015. SAE J2980: Considerations for ISO 26262 ASIL Hazard Classification.

Saldaña, J., 2016. The coding manual for qualitative researchers, Third. ed. SAGE.

Salmon, P.M., Cornelissen, M., Trotter, M.J., 2012. Systems-based accident analysis methods : A comparison of Accimap , HFACS , and STAMP. Saf. Sci. 50, 1158–1170. https://doi.org/10.1016/j.ssci.2011.11.009

Salmon, P.M., Lenné, M.G., Stanton, N.A., Jenkins, D.P., Walker, G.H., 2010. Managing error on the open road : The contribution of human error models and methods. Saf. Sci. 48, 1225–1235. https://doi.org/10.1016/j.ssci.2010.04.004

Sarter, N., 2008. Investigating mode errors on automated flight decks: illustrating the problem-driven, cumulative, and interdisciplinary nature of human factors research. Hum. Factors 50, 506–510. https://doi.org/10.1518/001872008X312233

Sauer, J., Nickel, P., Wastell, D., 2013. Designing automation for complex work environments under different levels of stress. Appl. Ergon. 44, 119–127. https://doi.org/10.1016/j.apergo.2012.05.008

Schöner, D.H., Neads, D.S., Schretter, N., 2009. Testing and Verification of Active Safety Systems with Coordinated Automated Driving, in: Proc. of the 21st International Technical Conference on Enhanced Safety of Vehicles (ESV), Stuttgart, Germany.

Schöner, H., Hurich, W., 2015. Testing with Coordinated Automated Vehicles, in: Handbook of Driver Assistance Systems. https://doi.org/10.1007/978-3-319-09840-1

Seppelt, B.D., Lee, J.D., 2007. Making adaptive cruise control (ACC) limits visible. Int. J. Hum. Comput. Stud.

65, 192–205. https://doi.org/10.1016/j.ijhcs.2006.10.001

Seppelt, B.D., Victor, T.W., 2016. Potential Solutions to Human Factors Challenges in Road Vehicle Automation. https://doi.org/10.1007/978-3-319-40503-2

Sepulcre, M., Gozalvez, J., Hernandez, J., 2013. Cooperative vehicle-to-vehicle active safety testing under challenging conditions. Transp. Res. PART C 26, 233–255. https://doi.org/10.1016/j.trc.2012.10.003

Shappell, S.A., Detwiler, C., Holcomb, K., Hackworth, C., Boquet, A., Wiegmann, D.A., 2007. Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. Hum. Factors 49, 227–242. https://doi.org/10.1518/001872007X312469

Sheridan, T.B., 1995. Human centered automation: oxymoron or common sense?, in: 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century. pp. 823–828. https://doi.org/10.1109/ICSMC.1995.537867

Shin, J., Bhat, C.R., You, D., Garikapati, V.M., Pendyala, R.M., 2015. Consumer preferences and willingness to pay for advanced vehicle technology options and fuel types. Transp. Res. Part C Emerg. Technol. 60, 511–524. https://doi.org/10.1016/j.trc.2015.10.003

Shladover, S.E., 2009. Cooperative (rather than autonomous) vehicle-highway automation systems. IEEE Intell. Transp. Syst. Mag. 1, 10–19. https://doi.org/10.1109/MITS.2009.932716

Singh, S., 2015. Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. (Traffic Safety Facts Crash Stats. Report No. DOT HS 812 115). Washington, DC.

SMMT, 2019. Connected and Autonomous Vehicles.

Son, J., Park, M., Park, B.B., 2015. The effect of age, gender and roadway environment on the acceptance and effectiveness of Advanced Driver Assistance Systems. Transp. Res. Part F Traffic Psychol. Behav. 31, 12–24. https://doi.org/10.1016/j.trf.2015.03.009

Sotomayor Martínez, R., 2015. System Theoretic Process Analysis of Electric Power Steering for Automotive. Massachusetts Institute of Technology.

Sparaco, P., 1995. Airbus seeks to keep pilot, new technology in harmony. Aviat. Week Space Technol. 62–63.

Sparrow, R., Howard, M., 2017. When human beings are like drunk robots : Driverless vehicles , ethics , and the future of transport. Transp. Res. Part C 80, 206–215. https://doi.org/10.1016/j.trc.2017.04.014

Stamatis, D.H., 2003. Failure mode and effect analysis : FMEA from theory to execution, 2nd ed. Milwaukee, Wisc. : ASQ Quality Press, 2003.

Stanton, N.A., Dunoyer, A., Leatherland, A., 2011. Detection of new in-path targets by drivers using Stop & Go Adaptive Cruise Control. Appl. Ergon. 42, 592–601. https://doi.org/10.1016/j.apergo.2010.08.016

Stanton, N.A., Salmon, P.M., Walker, G.H., Salas, E., Hancock, P.A., 2017. State-of-Science: Situation Awareness in individuals, teams and systems. Ergonomics 60, 1–33. https://doi.org/10.1080/00140139.2017.1278796

Stanton, N.A., Young, M., Mccaulder, B., 1997. Drive-by-Wire : The case of driver workload and reclaiming control with Adaptive Cruise Control. Saf. Sci. 27, 149–159.

Stanton, N.A., Young, M.S., 1998. Vehicle automation and driving performance. Ergonomics 41, 1014–1028. https://doi.org/10.1080/001401398186568

Stoop, J., Dekker, S., 2012. Are safety investigations pro-active ? Saf. Sci. 50, 1422–1430. https://doi.org/10.1016/j.ssci.2011.03.004

Strandberg, K., Olovsson, T., Jonsson, E., 2018. Securing the Connected Car: A Security-Enhancement Methodology. IEEE Veh. Technol. Mag. 13, 56–65. https://doi.org/10.1109/MVT.2017.2758179

Summala, H., 2000. Brake Reaction Times and Driver Behavior Analysis 2, 217–226.

Talebpour, A., Mahmassani, H.S., 2016. Influence of connected and autonomous vehicles on traffic flow stability and throughput. Transp. Res. Part C Emerg. Technol. 71, 143–163.

https://doi.org/10.1016/j.trc.2016.07.007

Tchiehe, D.N., Gauthier, F., 2017. Classification of risk acceptability and risk tolerability factors in occupational health and safety. Saf. Sci. 92, 138–147. https://doi.org/10.1016/j.ssci.2016.10.003

Thomas, J., Sgueglia, J., Suo, D., Leveson, N., Vernacchia, M., Sundaram, P., 2015. An Integrated Approach to Requirements Development and Hazard Analysis. https://doi.org/10.4271/2015-01-0274.Copyright

Tingvall, C., 1998. The Swedish "Vision Zero" and how parliamentary approval was obtained, in: Proc. of the Road Safety Research, Policing, Education Conference, Wellington, New Zealand. llington, New Zealand.

Tingvall, C., 1997. The Zero Vision: A Road Transport System Free from Serious Health Losses. Transp. Traffic Saf. Heal. New Mobil. 37–57.

Transport Systems Catapult, 2017. Taxonomy of Scenarios for Automated Driving.

Ulbrich, S., Menzel, T., Reschka, A., Schuldt, F., Maurer, M., 2015. Defining and Substantiating the Terms Scene , Situation , and Scenario for Automated Driving. https://doi.org/10.1109/ITSC.2015.164

Underwood, G., Crundall, D., Chapman, P., 2011. Driving simulator validation with hazard perception. Transp. Res. Part F Psychol. Behav. 14, 435–446. https://doi.org/10.1016/j.trf.2011.04.008

van Arem, B., Cornelie, J.G.V.D., Visser, R., 2005. The impact of Co-operative Adaptive Cruise Control on traffic flow characteristics. IEEE Trans. Intell. Transp. Syst. 7, 429–436.

Van Xanten, N.H.W., Pietersen, C.M., Pasman, H.J., Vrijling, H.K., Kerstens, J.G.M., 2013. Rituals in risk evaluation for land-use planning. Chem. Eng. Trans. 31, 85–90. https://doi.org/10.3303/CET1331015

Veland, H., Aven, T., 2015. Improving the risk assessments of critical operations to better reflect uncertainties and the unforeseen. Saf. Sci. 79, 206–212. https://doi.org/10.1016/j.ssci.2015.06.012

Verburg, D.J., Knaap, A.C.M. Van Der, Ploeg, J., 2002. VEHIL Developing and Testing Intelligent Vehicles, in: Proc. of the 2002 IEEE Intelligent Vehicles Symposium. Versailles.

Verma, M.K., Goertz, A.R., 2016. Preliminary Evaluation of Pre-Crash Safety System Effectiveness.

Vesely, W.E., Roberts, N.H., 1981. Fault Tree Handbook.

Villa, V., Paltrinieri, N., Khan, F., Cozzani, V., 2016. Towards dynamic risk analysis : A review of the risk assessment approach and its limitations in the chemical process industry. Saf. Sci. 89, 77–93. https://doi.org/10.1016/j.ssci.2016.06.002

W.P. Klockwork, 2012. Software on Wheels [WWW Document].

Wachenfeld, W., Winner, H., 2017a. The New Role of Road Testing for the Safety Validation of Automated Vehicles, in: Automated Driving. pp. 419–435. https://doi.org/10.1007/978-3-319-31895-0_17

Wachenfeld, W., Winner, H., 2017b. The New Role of Road Testing for the Safety Validation of Automated Vehicles. https://doi.org/10.1007/978-3-319-31895-0

Walker, G.H., Stanton, N.A., Salmon, P., 2016. Trust in vehicle technology 70, 157–182.

Weyuker, E.J., 1998. Testing component-based software: A cautionary tale. IEEE Softw. 15, 54–59. https://doi.org/10.1109/52.714817

Whittaker, J.A., 2000. What is software testing? And why is it so hard? IEEE Softw. 17, 70–79. https://doi.org/10.1109/52.819971

Wickens, C.D., 2008. Multiple Resources and Mental Workload 50, 449–455. https://doi.org/10.1518/001872008X288394.

Wiegmann, D., Shappell, S., 2001a. Human error analysis of commercial aviation accidents: Application of the Human Factors Analysis and Classification System (HFACS). Aviat. Space. Environ. Med.

Wiegmann, D., Shappell, S., 2001b. Applying the human factors analysis and classification system (HFACS) to the analysis of commercial aviation accident data, in: Proc. of the 11th International Symposium on Aviation Psychology. Columbus, Ohio.

Wiegmann, D.A., Shappell, S.A., 2001. A Human Error Analysis of Commercial Aviation Accidents Using the

Human Factors Analysis and Classification System (HFACS). Virginia, USA.

Wiener, E.L., Curry, R.E., 1980. Flight-deck automation: promises and problems. Ergonomics 23, 995–1011. https://doi.org/10.1017/CBO9781107415324.004

Winner, H., 2016. ADAS, Quo Vadis?, in: Handbook of Driver Assistance Systems.

Winter, J.C.F. De, Leeuwen, P.M. Van, Happee, R., 2012. Advantages and Disadvantages of Driving Simulators : A Discussion 2012, 47–50.

WMG, 2017. Drive-in, Driver-in-the-loop, multi-axis driving simulator (3xD) [WWW Document].

Wogalter, M.S., Brelsford, J.W., Desaulniers, D.R., Laughery, K.R., 1991. Consumer Product Warnings : The Role of Hazard Perception. J. Safety Res. 22, 71–82.

Woods, D.D., 1984. Visual momentum : a concept to improve the cognitive coupling of person and computer 229–244.

Xiong, H., Boyle, L.N., Moeckli, J., Dow, B.R., Brown, T.L., 2012. Use Patterns Among Early Adopters of Adaptive Cruise Control. Hum. Factors J. Hum. Factors Ergon. Soc. 54, 722–733. https://doi.org/10.1177/0018720811434512

Xu, J., Le, K., Deitermann, A., Montague, E., 2014. How different types of users develop trust in technology : A qualitative analysis of the antecedents of active and passive user trust in a shared technology. Appl. Ergon. 45, 1495–1503. https://doi.org/10.1016/j.apergo.2014.04.012

Young, K.L., Stephens, A.N., Logan, D.B., Lenn, M.G., 2017. Investigating the impact of static roadside advertising on drivers ' situation awareness 60, 136–145. https://doi.org/10.1016/j.apergo.2016.11.009

Young, M.S., Stanton, N. a, 2007. Back to the future: brake reaction times for manual and automated vehicles. Ergonomics 50, 46–58. https://doi.org/10.1080/00140130600980789

Young, M.S., Stanton, N.A., 2002. Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. Hum. Factors 44, 365–375. https://doi.org/10.1518/0018720024497709

Yu, H., Lin, C., Kim, B., 2016. Automotive Software Certification : Current Status and Challenges. https://doi.org/10.4271/2016-01-0050

Zeeb, K., Buchner, A., Schrauf, M., 2016. Is take-over time all that matters ? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. Accid. Anal. Prev. 92, 230–239. https://doi.org/10.1016/j.aap.2016.04.002

Zeeb, K., Buchner, A., Schrauf, M., 2015. What determines the take-over time? An integrated model approach of driver take-over after automated driving. Accid. Anal. Prev. 78, 212–221. https://doi.org/10.1016/j.aap.2015.02.023