Calibrating Trust through Knowledge: Introducing the Concept of Informed Safety for Automation in Vehicles

Siddartha Khastgir *, Stewart Birrell, Gunwant Dhadyalla, Paul Jennings
WMG, University of Warwick, UK
* Corresponding author: S.Khastgir.1@warwick.ac.uk

Abstract

There has been an increasing focus on the development of automation in vehicles due its many potential benefits like safety, improved traffic efficiency, reduced emissions etc. One of the key factors influencing public acceptance of automated vehicle technologies is their level of trust. Development of trust is a dynamic process and needs to be calibrated to the correct levels for safe deployment to ensure appropriate use of such systems. One of the factors influencing trust is the knowledge provided to the driver about the system's true capabilities and limitations. With a 56 participant driving simulator study, the authors found that with the introduction of knowledge about the true capabilities and limitations of the automated system, trust in the automated system increased as compared to when no knowledge was provided about the system. Participants experienced two different types of automated systems: low capability automated system and high capability automated system. Interestingly, with the introduction of knowledge, the average trust levels for both low and high capability automated systems were similar. Based on the experimental results, the authors introduce the concept of *informed safety*, i.e., informing the drivers about the safety limits of the automated system to enable them to calibrate their trust in the system to an appropriate level.

1. 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 etc.) in 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 (AD) features. The push towards ADAS and AD systems is driven due to their many potential benefits like increased safety leading to reducing the number of accidents (Tingvall, 1997; Guériau *et al.*, 2016; Cicchino, 2017), increased traffic throughput and road efficiency (Le Vine *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 (Stanton and Young, 1998; Balfe, Sharples and Wilson, 2015) and providing more productive time to drivers (Cairns *et al.*, 2014).

While it is important to provide drivers the opportunity to use ADAS and AD systems (with development in technology), it is equally important to ensure that the drivers actually use the systems in order to ensure the potential benefits from the use of such systems are realized (Lee and See, 2004; Diels and Bos, 2016). Unfortunately, the usage of ADAS features like ACC and Lane Departure Warning has been low (51% of highway driving time (Eichelberger and McCartt, 2014)). 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.

In the driving context, design and behaviour of ADAS and AD systems should be communicated to the driver (Stanton, Young and Mccaulder, 1997) and should be more human-like as it would make the driver-automation cooperation more transparent (Bifulco *et al.*, 2013; Casner, Hutchins and Norman, 2016; Wang *et al.*, 2016), leading to increased trust in the system. One of the challenges with the design of ADAS and AD is that their introduction changes drivers' task from active engagement to passive monitoring (van den Beukel, van der Voort and Eger, 2016). Drivers' driving task is said to have three different levels: 1) strategic 2) tactical and 3) operational (Michon, 1985). ADAS and AD systems alter these levels of driving tasks and the decision to design automation into any of the three levels is generally a trade-off decision (Johansson and Nilsson, 2016; Khastgir, Sivencrona, Dhadyalla, Billing, *et al.*, 2017). The trade-off decision determines the level of engagement of the driver in the driving task. The shift from active engagement to passive monitoring introduces new types of potential errors (human errors) in the driving task as the human driver is not suitable for the task of monitoring monotonous systems (Fitts *et al.*, 1951).

1.1. Trust

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). The designer of the automation makes assumptions about the best design for automation and distribution of driving tasks between the driver and the automated system. These assumption may or may not match with the drivers' perception of the automated system and task distribution. Muir (1994) has suggested that as the automation capability or reliability increases, trust also increases. 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, thus not benefiting from the system. Thus, in order to ensure appropriate use of the system, it is essential to calibrate drivers' trust to the appropriate level.

Trust is one of the most important factors influencing use of automation (Muir, 1987; Lee and Moray, 1992; Muir and Moray, 1996; Parasuraman and Riley, 1997; Parasuraman and Miller, 2004; Rudin-Brown and Parker, 2004; Walker, Stanton and Salmon, 2016). Before the authors discuss details of the development of trust, it is important to define trust in driving context. In order to define trust, the authors adapt 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; Xu *et al.*, 2014; Walker, Stanton and Salmon, 2016), with previous work conducted by the authors suggesting this can also include knowledge, certification, situation awareness, workload, self-confidence, experience, consequence and willingness (Khastgir, Birrell, Dhadyalla and Jennings, 2017). In this paper, authors discuss the effect of knowledge on trust.

1.1.1. Forms of trust

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 authors classify 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 paper.

1.1.2. 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, Heesen and Vollrath, 2013). In the absence of such knowledge, drivers may not be able to calibrate their trust to an appropriate level (Lee and See, 2004; Chavaillaz, Wastell and Sauer, 2016). While failures of automation has been proved to have a detrimental effect on trust, Lee and See (2004) argue that some failures can be classified as "good failures" with neglegible impact on trust. 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).

1.1.3. Types of knowledge

Based on literature (Rasmussen, 1985; Seppelt and Lee, 2007; Xu *et al.*, 2014; Biassoni, Ruscio and Ciceri, 2016; Feldhütter *et al.*, 2016; Miller *et al.*, 2016; Bennett, 2017), the following classification for knowledge about the capabilities and limitations of automated systems was proposed by (Khastgir, Birrell, Dhadyalla and Jennings, 2017):

- Static knowledge: Understanding of the functionality of the automated system (intentions behind the design of the system and functionality) (Larsson, 2012; Eichelberger and McCartt, 2014). 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.

- 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, 2016) and increase transparency in the cooperation between humans and automation (Eriksson and Stanton, 2017). While in-vehicle information systems (IVISs) are known to have detrimental effect on driving performance (Peng, Boyle and Lee, 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.

- 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, Hutchins and Norman, 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 authors suggest 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).

1.1.4. 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, Gozalvez and Hernandez, 2013). As discussed in section 1.1.2, 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 AD systems faces reliability challenges (Khastgir, Birrell, Dhadyalla, Sivencrona, *et al.*, 2017) and validation challenges which include challenges in test methods and test setup (Hendriks, Pelders and Tideman, 2010; Khastgir *et al.*, 2015; Yu, Lin and Kim, 2016). While the authors consider knowledge creation as an important part of the process of development of trust, it remains out of scope of this paper and will be discussed in future publications.

While defining trust in section 1.1, the authors mentioned that trust is a history dependent construct, suggesting its dynamic nature. The authors adopt the definition of calibration of trust as *"the process of adjusting trust to correspond to an objective measure of trustworthiness"* (Muir, 1994). Khastgir et al. (2017a) introduced 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). There can be various intervention methods to potentially increase/adjust trust in different stages of calibration. In this paper, the authors discuss the use of static knowledge as an intervention method in the process of calibration of trust.

1.2. Research Question

As discussed in section 1.1, many authors have studied the effect of reliability (or automation capability) on trust (Muir, 1994; Muir and Moray, 1996; Chavaillaz, Wastell and Sauer, 2016), 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 a driving simulator study, this paper aims to answer the following two research questions:

1. Does providing static knowledge about the automation capability of the system influence *"trust in the system"*?
2. With static knowledge about the automation capability, does automation capability influence *"trust in the system"*?

1.2.1. Hypothesis

The authors hypothesize that static knowledge influences *"trust in the system"* as it would help influence drivers' mental model and aid in them exercising their knowledge-based behaviour in unfamiliar situations. Furthermore, the authors believe that static knowledge would have limited effect on drivers' *"trust with the system"* as drivers' lack information about the automation health and its intentions. While static knowledge does provide an ability for drivers to predict failures, it does not help them understand the real-time tactical and operational driving task choices made by the automated system.

This paper is organized in five sections. Section two discusses the methodology adopted for the study, section three illustrates the results of the study, section four provides a discussion on the results and the paper concludes with a conclusion in section five.

2. Methodology

2.1. 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 360° 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.

2.2. 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 2.3.2. 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 1).

Table 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

2.3. 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, 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 2). Both these factors have been discussed in section 1.1 as being essential to evaluate development of trust. 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, Banks and Stanton, 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 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

2.3.1. Hazards

In order to choose the five hazardous events, a hazard analysis of an automated vehicle was conducted as per the ISO 26262 (ISO, 2011) functional safety process. Five different automated vehicle functions were identified and a hazard was identified for each of the functions (Table 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 3: Description of five hazardous events

Function	Hazard	Hazardous event description
Braking	Lack of Braking	Pedestrian suddenly changes direction and comes in front of the 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.

2.3.2. 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 2.3.1. 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.

2.4. 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 a researcher 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.

At the end of each experiment run, participants were asked to fill a trust rating questionnaire (section 2.4.2), Simulator Sickness Questionnaire (SSQ) (Kennedy *et al.*, 1993), and Van Der Laan's acceptance questionnaire (Van Der Laan, Heino and De Waard, 1997). However, the results from the latter two haven't been included in this paper.

2.4.1. Imparting 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 two scripts are examples of the how knowledge was imparted to the participants.

Example 1: "The automated control system from the supplier is 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 1). In your drive today, you might have witnessed bright 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, 2017) where the white rear end of the truck was not detected against the bright sky."

Figure 1: Camera view while driving in fog
(image source: https://www.flickr.com/photos/kubina/2160242894; date accessed: 2017-12-04)

Example 2: "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 2). 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 2). 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."

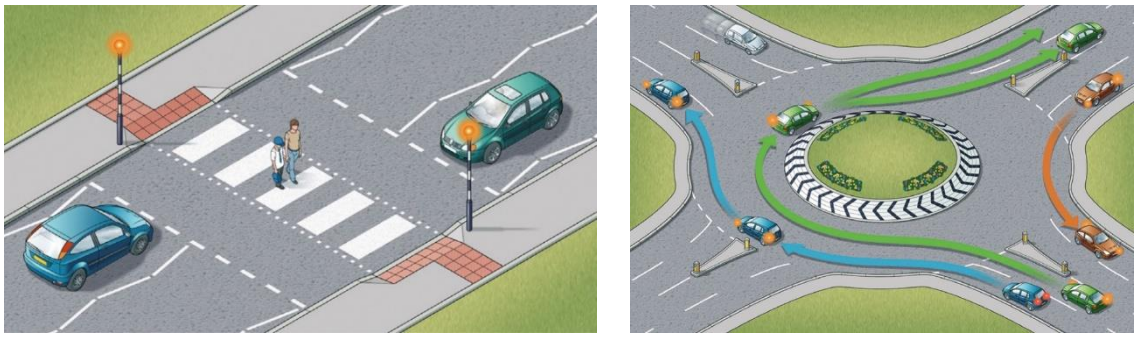

Figure 2: Rules of road: rule 19 (left) and rule 185(right). (DfT, 2017)

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 Figure 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.

2.4.2. Trust questionnaire

At the end of each of the two experiment runs, participants were asked to rate their level of *"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, Anceaux and Vienne, 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, Bisantz and Drury, 2000), couldn't be used as they don't classify trust into the two components mentioned in section 1.1. 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 in 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"*

On the trust scale, 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.

3. Results

3.1. 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 3). 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 3). 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 4). 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 3).

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. 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), 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, $F(1, 13) = 0.020$, $p = 0.89$ ($\eta_p^2 = 0.002$).

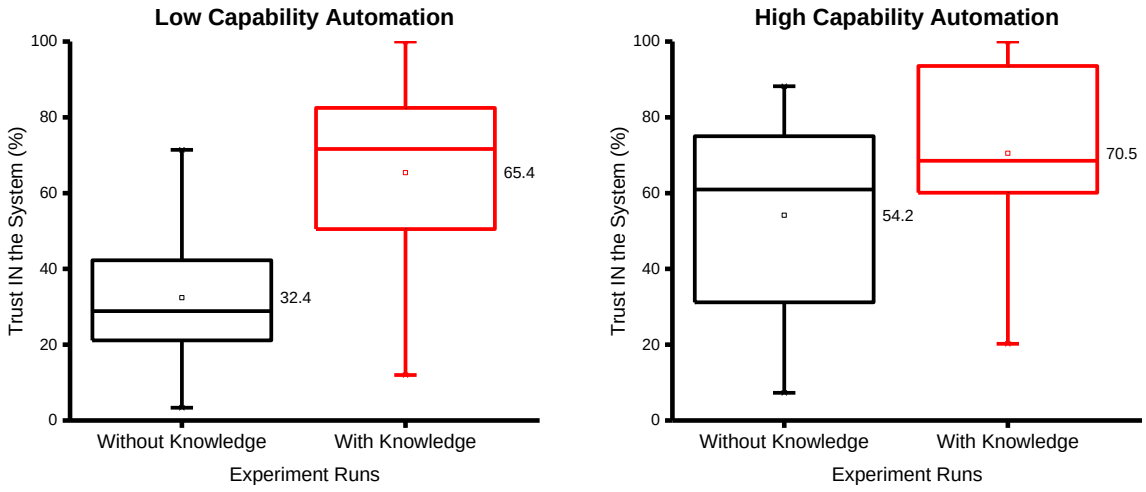

Figure 3: Box-plots of Trust-In the system ratings (highlighting average trust ratings) (central dot represents average value)

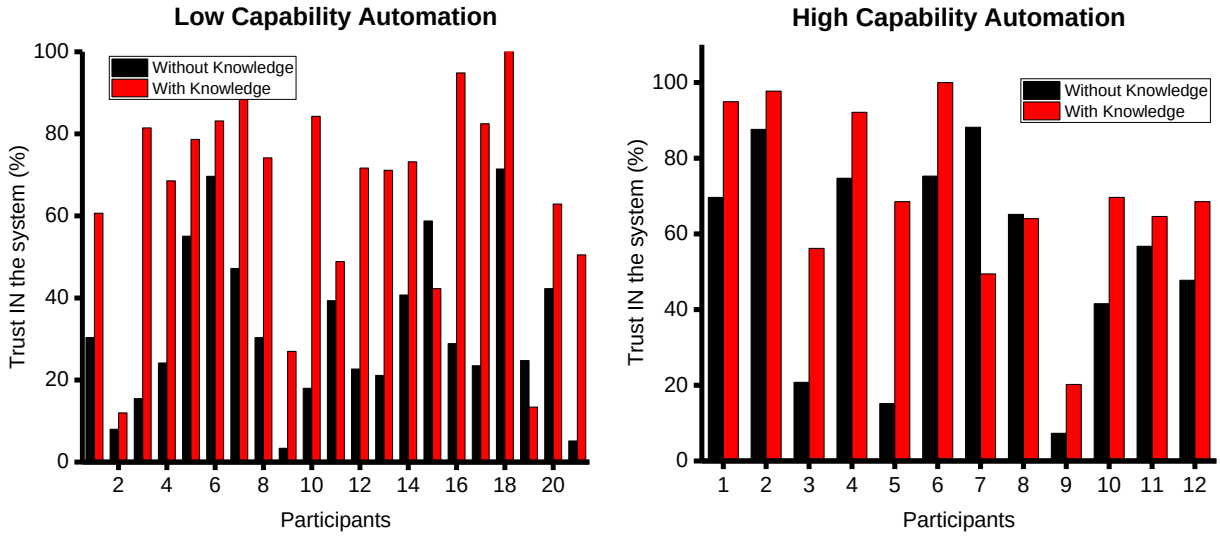

Figure 4: "Trust in the System" level of individual participants for low capability and high capability automation

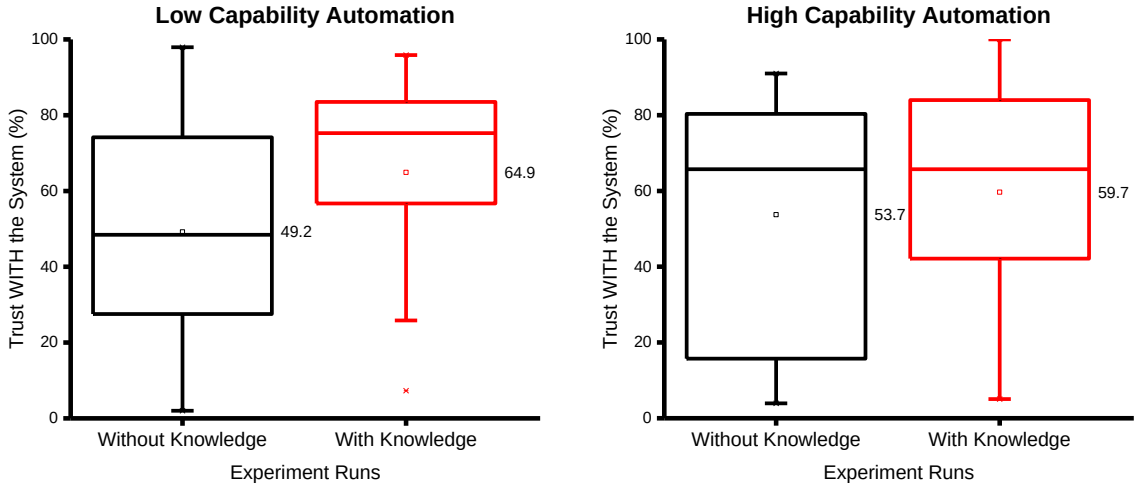

Figure 5: Box-plots of Trust-With the system ratings (central dot represents average value)

3.2. 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. The authors define a false press 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 6). 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.

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 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.

As discussed in section 2.4.1, for the low capability automation group, participants were given a lot of knowledge based on the automated systems' 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 4).

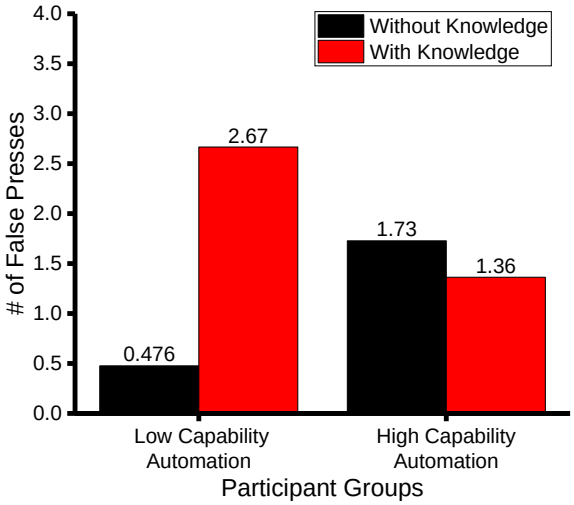

Figure 6: Average number of false presses

3.3. Accidents

The authors define an accident 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 7). For high capability automation, the average number of accidents reduced slightly from 0.58 to 0.42 (Figure 7). 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.

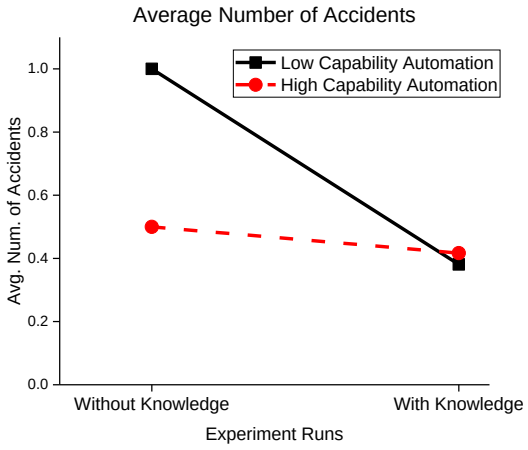

Figure 7: Average number of accidents

4. Discussion

As mentioned in section 1.1.1, *"trust in the system"* refers to the capability of the system where as *"trust with the system"* refers to the ability of the driver to work with the system. In the study presented, the authors have illustrated that with the introduction of 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, Lorenz and Krems, 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 1.1.2, 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"*.

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, 2017). Thus, the authors suggest that *"trust with the system"* is potentially influenced to a larger extent by dynamic (real-time) knowledge about the system capabilities and limitation.

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.

4.1. Informed Safety

Results from this study could infer that vehicle manufacturers may choose to introduce low-capability systems and provide 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 knowledge increased the level of trust in the system significantly (from 32.4% to 65.4%), it also increased the number of false presses significantly (from 0.476 to 2.67). Therefore, very low capability and too much knowledge is also not an appropriate solution. The authors believe that there is an optimum level of system capability and knowledge to be imparted at which trust could be maximized and false presses 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 trust in the system.

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 this study 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.

This paper 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 confidence 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.

4.2. Future research

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, Kircher and Hultgren, 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 of 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.

4.3. Study limitations

The WMG's 3xD simulator provides a fully immersive driving experience for participants. However, like all simulator studies, transferability of results to real world needs to be evaluated separately. Real-world evaluation of trust remains out of the scope of this paper. Additionally, as discussed in section 4.1, informed safety, as introduced in this paper, has two facets: 1) static knowledge (e.g. initial briefing and driving manual) and 2) dynamic knowledge such as human-machine interface. In this paper, the authors only provided static informed safety to drivers. Future studies are planned where participants will be provided both dynamic knowledge and static knowledge. Results will be published in future publications.

5. Conclusion

Trust in automated systems is one of the key factors that would help realize the potential benefits offered by the introduction of automation in vehicles. However, trust level needs to be calibrated to the appropriate level in order to reap the benefits of the automated systems in a safe manner by preventing misuse or disuse. This study explores the effect of knowledge about the automation capability on trust in the system.

In this paper, the authors demonstrate via a 56 participants driving simulator study that "trust in the system" increases with the introduction of static knowledge about the capabilities and limitation of the automated system. With the introduction of static knowledge, trust in the system for both low capability automation and high capability automation were not significantly different, 65.4% and 70.5% respectively, suggesting no influence of automation capability on trust in the system when knowledge is provided to the drivers. Based on results, the authors introduced the concept of *"informed safety"* which helps calibrate drivers' trust to an appropriate level, subsequently ensuring safe use of the automated system.

Interestingly, with the introduction of static knowledge the average number of false presses had contrasting results for the two automation groups. With the introduction of knowledge, for the high capability automation group, the average number of false presses decreased from 1.73 to 1.36, while it increased from 0.47 to 2.67 for the low capability automation group. However, average number of accidents decreased from 1 to 0.38 and from 0.58 to 0.42 for low capability automation and high capability automation respectively. The improved safety with the introduction knowledge lends its support to the concept of informed safety. In order to reduce the number of false presses, the authors hypothesize the need to provide "informed safety" in a dynamic manner, i.e., via knowledge about the automation state and health through the HMI system. Results on the study exploring the hypothesis will be presented in future publications.

Acknowledgements

The work presented in this paper has been carried under the EPSRC Grant (Grant EP/K011618/1). The authors would like to thank the WMG centre of HVM Catapult and WMG, University of Warwick, UK, for providing the necessary infrastructure for conducting this study. WMG hosts one of the seven centres that together comprise the High Value Manufacturing Catapult in the UK. The authors would also like to thank Andrew D. Moore and Jonathan Smith for their assistance in building the experimental setup. The authors would also like to thank three anonymous reviewers for their detailed comments on previous versions of the paper, which has helped considerably to improve the paper.

References

Bainbridge, L. (1983) 'Ironies of automation', *Automatica*, 19(6), pp. 775–779. doi: 10.1016/0005-1098(83)90046-8.
Balfe, N., Sharples, S. and Wilson, J. R. (2015) 'Impact of automation: Measurement of performance, workload and behaviour in a complex control environment', *Applied Ergonomics*. Elsevier Ltd, 47, pp. 52–64. doi: 10.1016/j.apergo.2014.08.002.
Banks, V. A. and Stanton, N. A. (2016) 'Keep the driver in control: Automating automobiles of the future', *Applied Ergonomics*. Elsevier Ltd, 53, pp. 389–395. doi: 10.1016/j.apergo.2015.06.020.
Beggiato, M. and Krems, J. F. (2013) 'The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information', *Transportation Research Part F: Traffic Psychology and Behaviour*, 18, pp. 47–57. doi: 10.1016/j.trf.2012.12.006.
Beller, J., Heesen, M. and Vollrath, M. (2013) 'Improving the Driver-Automation Interaction: An Approach Using Automation Uncertainty', *Human Factors*, 55(6), pp. 1130–1141. doi: 10.1177/0018720813482327.
Bennett, K. B. (2017) 'Ecological interface design and system safety: One facet of Rasmussen's legacy', *Applied Ergonomics*. Elsevier Ltd, 59, pp. 625–636. doi: 10.1016/j.apergo.2015.08.001.
van den Beukel, A. P., van der Voort, M. C. and Eger, A. O. (2016) 'Supporting the changing driver's task: Exploration of interface designs for supervision and intervention in automated driving', *Transportation Research Part F: Traffic Psychology and Behaviour*, 43, pp. 279–301. doi: 10.1016/j.trf.2016.09.009.
Biassoni, F., Ruscio, D. and Ciceri, R. (2016) 'Limitations and automation: The role of information about device-specific features in ADAS acceptability', *Safety Science*, 85, pp. 179–186. doi: 10.1016/j.ssci.2016.01.017.
Bifulco, G. N. *et al.* (2013) 'Driving behaviour models enabling the simulation of Advanced Driving Assistance Systems: Revisiting the Action Point paradigm', *Transportation Research Part C: Emerging Technologies*, 36, pp. 352–366. doi: 10.1016/j.trc.2013.09.009.
Cairns, S. *et al.* (2014) 'Sociological perspectives on travel and mobilities: A review', *Transportation Research Part A: Policy and Practice*, 63, pp. 107–117. doi: 10.1016/j.tra.2014.01.010.
Casner, S. M., Hutchins, E. L. and Norman, D. (2016) 'The Challenges of Partially Automated Driving', *Communications of the ACM*, 59(5), pp. 70–77. doi: 10.1145/2830565.
Chavaillaz, A., Wastell, D. and Sauer, J. (2016) 'System reliability, performance and trust in adaptable automation', *Applied Ergonomics*, 52, pp. 333–342. doi: 10.1016/j.apergo.2015.07.012.
Cicchino, J. B. (2017) *Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates*, *Accident Analysis and Prevention*. doi: 10.1016/j.aap.2016.11.009.
DfT (2017) *The Highway Code*. Available at: https://www.gov.uk/guidance/the-highway-code (Accessed: 18 July 2017).
Diels, C. and Bos, J. E. (2016) 'Self-driving carsickness', *Applied Ergonomics*, 53, pp. 374–382. doi: 10.1016/j.apergo.2015.09.009.
Eichelberger, A. H. and McCartt, A. T. (2014) 'Volvo drivers' experiences with advanced crash avoidance and related technologies.', *Traffic Injury Prevention*, 15(2), pp. 187–195. doi: 10.1080/15389588.2013.798409.
Eriksson, A., Banks, V. A. and Stanton, N. A. (2017) 'Transition to manual: Comparing simulator with on-road control transitions', *Accident Analysis and Prevention*, 102, pp. 227–234. doi: 10.1016/j.aap.2017.03.011.
Eriksson, A. and Stanton, N. A. (2017) 'The chatty co-driver: A linguistics approach applying lessons learnt from aviation incidents', *Safety Science*, 99, pp. 94–101. doi: 10.1016/j.ssci.2017.05.005.
Fagnant, D. J. and Kockelman, K. (2015) 'Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations', *Transportation Research Part A: Policy and Practice*, 77, pp. 167–181. doi: 10.1016/j.tra.2015.04.003.
Fagnant, D. J. and Kockelman, K. M. (2014) 'The travel and environmental implications of shared autonomous vehicles, using agent-based

model scenarios', *Transportation Research Part C: Emerging Technologies*, 40, pp. 1–13. doi: 10.1016/j.trc.2013.12.001.
Feldhütter, A. *et al.* (2016) 'Trust in Automation as a matter of media and experience of automated vehicles.', in *Proc. of the Human Factors and Ergonomics Society 60th Annual Meeting*, pp. 2024–2028.
Fitts, P. M. *et al.* (1951) *Human engineering for an effective air - navigation and traffic - control system*. Washington, D.C., USA.
Guériau, M. *et al.* (2016) 'How to assess the benefits of connected vehicles? A simulation framework for the design of cooperative traffic management strategies', *Transportation Research Part C: Emerging Technologies*, 67, pp. 266–279. doi: 10.1016/j.trc.2016.01.020.
Helldin, T. *et al.* (2013) 'Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving', in *Proc. of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13*, pp. 210–217. doi: 10.1145/2516540.2516554.
Hendriks, F., Pelders, R. and Tideman, M. (2010) 'Future Testing of Active Safety Systems', *SAE International Journal of Passenger Cars - Electronic and Electrical Systems*, 3(2), pp. 170–175. doi: 10.4271/2010-01-2334.
Hergeth, S., Lorenz, L. and Krems, J. F. (2017) 'Prior Familiarization With Takeover Requests Affects Drivers' Takeover Performance and Automation Trust', *Human Factors*, 59(3), pp. 457–470. doi: 10.1177/0018720816678714.
ISO (2011) *Road vehicles — Functional safety (ISO 26262)*.
Jian, J.-Y., Bisantz, A. M. and Drury, C. G. (2000) 'Foundations for an Empirically Determined Scale of Trust in Automated System', *International Journal of Cognitive Ergonomics*, 4(1), pp. 53–71.
Johansson, R. and Nilsson, J. (2016) 'The Need for an Environment Perception Block to Address all ASIL Levels Simultaneously', in *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*. Gothenburg, Sweden. doi: 10.1109/IVS.2016.7535354.
Kennedy, R. S. *et al.* (1993) 'Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness', *International Journal of Aviation Psychology*, 3(3), pp. 203–220.
Khastgir, S. *et al.* (2015) 'Identifying a Gap in Existing Validation Methodologies for Intelligent Automotive Systems : Introducing the 3xD Simulator', in *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*. Seoul, South Korea: IEEE, pp. 648–653. doi: 10.1109/IVS.2015.7225758.
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. doi: 10.1007/978-3-319-41682-3_45.
Khastgir, S., Sivencrona, H., Dhadyalla, G., Billing, P., *et al.* (2017) 'Introducing ASIL Inspired Dynamic Tactical Safety Decision Framework for Automated Vehicles', in *Proc. of the IEEE 20th International Conference on Intelligent Transportation Systems (ITSC 2017)*. Yokohama, Japan, pp. 2398–2403.
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.
Van Der Laan, J. D., Heino, A. and De Waard, D. (1997) 'A simple procedure for the assessment of acceptance of advanced transport telematics', *Transportation Research Part C: Emerging Technologies*, 5(1), pp. 1–10. doi: 10.1016/S0968-090X(96)00025-3.
Larsson, A. F. L. (2012) 'Driver usage and understanding of adaptive cruise control', *Applied Ergonomics*, 43, pp. 501–506. doi: 10.1016/j.apergo.2011.08.005.
Larsson, A. F. L., Kircher, K. and Hultgren, J. A. (2014) 'Learning from experience: Familiarity with ACC and responding to a cut-in situation in automated driving', *Transportation Research Part F: Traffic Psychology and Behaviour*, 27, pp. 229–237. doi: 10.1016/j.trf.2014.05.008.
Lee, J. D. and See, K. A. (2004) 'Trust in Automation: Designing for Appropriate Reliance', *Human factors*, 46(1), pp. 50–80. doi: 10.1518/hfes.46.1.50.30392.
Lee, J. and Moray, N. (1992) 'Trust, control strategies and allocation of function in human-machine systems', *Ergonomics*, 35(10), pp. 1243–1270. doi: 10.1080/00140139208967392.
Louw, T. and Merat, N. (2017) 'Are you in the loop? Using gaze dispersion to understand driver visual attention during vehicle automation', *Transportation Research Part C*. Elsevier Ltd, 76, pp. 35–50. doi: 10.1016/j.trc.2017.01.001.
Michon, J. A. (1985) 'A critical view of driver behavior models: what do we know, what should we do?', in Evans, L. and Schwing, R. C. (eds) *Human behavior and traffic safety*. Plenum Press, pp. 485–520. doi: 10.1007/978-1-4613-2173-6.
Miller, D. *et al.* (2016) 'Behavioral Measurement of Trust in Automation: The Trust Fall', in *Proc. of the Human Factors and Ergonomics Society 2016 Annual Meeting*, pp. 1849–1853. doi: 10.1177/1541931213601422.
Molesworth, B. R. C. and Koo, T. T. R. (2016) 'The influence of attitude towards individuals??? choice for a remotely piloted commercial flight: A latent class logit approach', *Transportation Research Part C: Emerging Technologies*, 71, pp. 51–62. doi: 10.1016/j.trc.2016.06.017.
Muir, B. M. (1987) 'Trust between humans and machines, and the design of decision aids', *International Journal of Man-Machine Studies*, 27, pp. 527–539. doi: 10.1016/S0020-7373(87)80013-5.
Muir, B. M. (1994) 'Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems', *Ergonomics*, 37(11), pp. 1905–1922. doi: 10.1080/00140139408964957.
Muir, B. M. and Moray, N. (1996) 'Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation.', *Ergonomics*, 39(3), pp. 429–460. doi: 10.1080/00140139608964474.
NHTSA (2017) *Investigation Report: PE 16-007 (MY2014-2016 Tesla Model S and Model X)*.
Parasuraman, R. and Miller, C. a. (2004) 'Trust and etiquette in high-criticality automated systems', *Communications of the ACM*, 47(4), pp. 51–55. doi: 10.1145/975817.975844.
Parasuraman, R. and Riley, V. (1997) 'Humans and Automation: Use, Misuse, Disuse, Abuse', *Human Factors*, 39(2), pp. 230–253.
Peng, Y., Boyle, L. N. and Lee, J. D. (2014) 'Reading, typing, and driving: How interactions with in-vehicle systems degrade driving performance', *Transportation Research Part F: Traffic Psychology and Behaviour*, 27, pp. 182–191. doi: 10.1016/j.trf.2014.06.001.
Rajaonah, B. *et al.* (2008) 'The role of intervening variables in driver-ACC cooperation', *International Journal of Human Computer Studies*, 66(3), pp. 185–197. doi: 10.1016/j.ijhcs.2007.09.002.
Rajaonah, B., Anceaux, F. and Vienne, F. (2006) 'Trust and the use of adaptive cruise control: a study of a cut-in situation', *Cognition, Technology & Work*, 8(2), pp. 146–155. doi: 10.1007/s10111-006-0030-3.
Rasmussen, J. (1983) 'Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models', *IEEE Transactions on Systems, Man, and Cybernetics*, 13(3), pp. 257–266.
Rasmussen, J. (1985) 'The Role of Hierarchical Knowledge Representation in Decisionmaking and System Management', *IEEE Transactions on Systems, Man, and Cybernetics*, 15(2), pp. 234–243. doi: 10.1109/TSMC.1985.6313353.
Rudin-Brown, C. M. and Parker, H. a. (2004) 'Behavioural adaptation to adaptive cruise control (ACC): Implications for preventive strategies', *Transportation Research Part F: Traffic Psychology and Behaviour*, 7(2), pp. 59–76. doi: 10.1016/j.trf.2004.02.001.
SAE (2018) *Surface Vehicle Recommended Practice, J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems*

for On-Road Motor Vehicles. doi: 10.4271/2012-01-0107.
Seppelt, B. D. and Lee, J. D. (2007) 'Making adaptive cruise control (ACC) limits visible', *International Journal of Human Computer Studies*, 65(3), pp. 192–205. doi: 10.1016/j.ijhcs.2006.10.001.
Sepulcre, M., Gozalvez, J. and Hernandez, J. (2013) 'Cooperative vehicle-to-vehicle active safety testing under challenging conditions', *Transportation Research Part C: Emerging Technologies*, 26, pp. 233–255. doi: 10.1016/j.trc.2012.10.003.
Stanton, N. A. and Young, M. S. (1998) 'Vehicle automation and driving performance', *Ergonomics*, 41(7), pp. 1014–1028. doi: 10.1080/001401398186568.
Stanton, N. a, Young, M. and Mccaulder, B. (1997) 'Drive-By-Wire : the Case of Driver Workload and Reclaiming Control With Adaptive Cruise Control', *Safety Science*, 27(2), pp. 149–159. doi: 10.1016/S0925-7535(97)00054-4.
Talebpour, A. and Mahmassani, H. S. (2016) 'Influence of connected and autonomous vehicles on traffic flow stability and throughput', *Transportation Research Part C: Emerging Technologies*, 71, pp. 143–163. doi: 10.1016/j.trc.2016.07.007.
Tingvall, C. (1997) 'The Zero Vision: A Road Transport System Free from Serious Health Losses', *Transportation, Traffic Safety and Health: the New Mobility*, pp. 37–57.
Le Vine, S. *et al.* (2016) 'Automated cars: Queue discharge at signalized intersections with "Assured-Clear-Distance-Ahead" driving strategies', *Transportation Research Part C: Emerging Technologies*, 62, pp. 35–54. doi: 10.1016/j.trc.2015.11.005.
Walker, G. H., Stanton, N. A. and Salmon, P. (2016) 'Trust in Vehicle Technology', *International Journal of Vehicle Design*, 70(2), pp. 157–182. doi: 10.1504/IJVD.2016.074419.
Wang, J. *et al.* (2016) 'Driving safety field theory modeling and its application in pre-collision warning system', *Transportation Research Part C: Emerging Technologies*, 72, pp. 306–324. doi: 10.1016/j.trc.2016.10.003.
WMG (2017) *Drive-in, Driver-in-the-loop, multi-axis driving simulator (3xD)*. Available at: http://www2.warwick.ac.uk/fac/sci/wmg/research/naic/facilities/ (Accessed: 10 July 2017).
Xu, J. *et al.* (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', *Applied Ergonomics*, 45(6), pp. 1495–1503. doi: 10.1016/j.apergo.2014.04.012.
Yu, H., Lin, C.-W. and Kim, B. (2016) 'Automotive Software Certification: Current Status and Challenges', *SAE International Journal of Passenger Cars - Electronic and Electrical Systems*, 9(1), pp. 74–80. doi: 10.4271/2016-01-0050.