Improving Collaboration Between Drivers and Automated Vehicles with Trust Processing Methods

by

Hebert Azevedo Sá

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Robotics)
in The University of Michigan
2021

Doctoral Committee:

Associate Professor Lionel Robert, Co-Chair
Professor Dawn Tilbury, Co-Chair
Professor Brent Gillespie
Professor Nadine Sarter
Assistant Professor X. Jessie Yang

Hebert Azevedo Sá

azevedo@umich.edu

ORCID iD: 0000-0001-7301-6685

ACKNOWLEDGEMENTS

I must express my most sincere gratitude to all organizations that, through their support and resources, allowed me to pursue my doctoral degree and made this work possible: the Brazilian Army's Military Institute of Engineering and Department of Science and Technology; the Robotics Institute, the Rackham Graduate School and the University of Michigan; the Automotive Research Center (ARC) and the United States Army Combat Capabilities Development Command/Ground Vehicle Systems Center (GVSC).

I am grateful for the friendship and support from my MAVRIC labmates Huajing Zhao, Qiaoning Zhang, Na Du, Connor Esterwood, and, more recently, Arsha Ali. In particular, I would like to especially thank my friend Suresh Kumaar Jayaraman, who was always willing to guide me in my first steps as a Ph.D. student. I also must thank Prof. Kira Barton's students who shared their lab and study spaces with us, and all Robotics Ph.D. students who took the course-based qualifying exam by the end of the Winter/2019 term. Speaking of the Robotics students, I could never forget to thank our Graduate Program Coordinator Denise Edmund for all her support since before I came to Ann Arbor.

I could not thank my advisors, Professor Lionel Robert and Professor Dawn Tilbury, enough. They were very sensitive to my particular needs and my goal of finishing the doctoral program earlier than usual. They gave me a lot of support in these three intense years and did their best to teach me how to do high-quality research. I also would like to thank Professor X. Jessie Yang, particularly for her

help with human factors engineering topics, and the other members of my committee, Professor Nadine Sarter and Professor Brent Gillespie, for their helpful feedback early on my research.

Most importantly, I am thankful for my wonderful family. My parents, for all their endless love and support. My sister, for her friendship and for being the only child around for the past years. My son, a gift from God, and the most fantastic woman I have met in my life—who happens to be his mother and my lovely wife.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Trust has gained attention in the Human-Robot Interaction (HRI) field, as it is considered an antecedent of people's reliance on machines. In general, people are likely to rely on and use machines they trust, and to refrain from using machines they do not trust. Recent advances in robotic perception technologies open paths for the development of machines that can be aware of people's trust by observing their human behaviors. This dissertation explores the role of trust in the interactions between humans and robots, particularly Automated Vehicles (AVs). Novel methods and models are proposed for perceiving and processing drivers' trust in AVs and for determining both humans' natural trust and robots' artificial trust.

Two high-level problems are addressed in this dissertation: (1) the problem of avoiding or reducing miscalibrations of drivers' trust in AVs, and (2) the problem of how trust can be used to dynamically allocate tasks between a human and a robot that collaborate.

A complete solution is proposed for the problem of avoiding or reducing trust miscalibrations. This solution combines methods for estimating and influencing drivers' trust through interactions with the AV. Three main contributions stem from that solution: (i) the characterization of risk factors that affect drivers' trust in AVs, which provided theoretical evidence for the development of a linear model for driver trust in AVs; (ii) the development of a new method for real-time trust estimation, which leveraged the trust linear model mentioned above for the implementation of a Kalman-filter-based approach, able to provide numerical estimates from the processing of drivers' behavioral measurements; and (iii) the development of a new method

for trust calibration, which identifies trust miscalibration instances from comparisons between drivers' trust in the AV and that AV's capabilities, and triggers messages from the AV to the driver. These messages are effective for encouraging or warning drivers that are undertrusting or overtrusting the AV capabilities respectively as shown by the obtained results.

Although the development of a trust-based solution for dynamically allocating tasks between a human and a robot (i.e., the second high-level problem addressed in this dissertation) remains an open problem, we take a step forward in that direction. The fourth contribution of this dissertation is the development of a unified bi-directional model for predicting natural and artificial trust. This trust model is based on mathematical representations of both the trustee agent's capabilities and the required capabilities for the execution of a task. Trust emerges from comparisons between the agent capabilities and task requirements, roughly replicating the following logic: if a trustee agent's capabilities exceed the requirements for executing a certain task, then the agent can be highly trusted (to execute that task); conversely, if that trustee agent's capabilities fall short of that task requirements, trust should be low. In this trust model, the agent's capabilities are represented by random variables that are dynamically updated over interactions between the trustor and the trustee whenever the trustee is successful or fails in the execution of a task. These capability representations allow for the numerical computation of human's trust or robot's trust, which is represented by the probability of a given trustee agent to execute a given task successfully.

CHAPTER I

Introduction

1.1 Motivation

Trust is a topic that has recently received considerable attention from human-robot interaction (HRI) researchers [64]. Trust facilitates cooperation between people, as well as between people and automated systems, like robots. [44]. HRI researchers expect that, in the future, robots will be able to understand people's behaviors and adapt their own behaviors to enable seamless human-robot interactions. Robots will likely have to take people's trust into consideration when autonomously making decisions or taking physical actions in their operating environment [36].

The intent of this dissertation is to extend the state-of-the-art trust-related knowledge, in order to solve problems that emerge when people interact with robots—particularly when these robots are Automated Vehicles (AVs) and those people assume the role of AV drivers. To accomplish this, this dissertation proposes methods for trust processing—i.e., measuring and influencing trust—that are useful for solving two main research problems: reducing trust miscalibrations and dynamically allocating tasks between human and robot collaborators.

Trust miscalibrations occur when there is a misalignment between the human operators' trust in the system and the system's capabilities [60]. Trust miscalibrations are likely to lead to inappropriate reliance on a system. A solution offered in this

dissertation to reducing trust miscalibrations is to implement a trust estimator and a trust calibrator that are able to manage trust.

Dynamic task allocation refers to assigning tasks to the human operator and to the robot, considering their different capabilities when they work together in a team. Establishing an analogy between human-human teams and human-robot teams, trust is a key element of the task allocation problem when agents are peers. Each agent has its own opinion of which agent should be executing each task, and this opinion should be based on trust.

The following chapters will explore how trust between drivers and AVs can be processed to solve the problems described above in the driver-AV interaction context. Each chapter presents a contribution and is based on a journal article that has already been published. Those articles are directed at answering the following four high-level research questions:

(i) how does risk affect AV trust and drivers' trusting behaviors (addressed in Chapter III, which is based on publication [10])?

(ii) how to measure drivers' trust in AVs (addressed in Chapter IV, based on publication [6])?

(iii) how to influence and calibrate drivers' trust in AVs (addressed in Chapter V, based on publication [7])?

(iv) how to use trust to assign tasks between a human and an automated system, which, in this case, could be an AV or a different robot (addressed in Chapter VI, based on publications [8] and in [9])?

1.1.1 How Does Risk Affect AV Trust and Drivers' Trusting Behaviors?

The Society of Automotive Engineers (SAE) defines six levels of driving automation ranging from 0 to 5, where 0 stands for "no driving automation" and 5 for "full

Figure 1.1: SAE J3016 levels of driving automation [101].

driving automation". Figure 1.1 presents a summarized description of these automation levels and their main characteristics. The term *automated driving system* (ADS) is used to describe driving automation systems that can be classified at levels 3, 4, and 5. The ADS consists of the hardware and software elements that distinguish AVs from human-driven vehicles. ADSs are automotive technologies that allow vehicles to engage in self-driving under specific conditions without any input from the driver [68]. An important benefit of such a system is the potential for drivers to engage in non-driving-related tasks (NDRTs) such as online activities (e.g., texting, checking their email, or surfing the web) [24, 25, 32, 71]. Along with the potential safety benefits, the increase in productivity through NDRTs is often cited as a relevant motivation behind the adoption of ADSs [35, 88, 94].

Trust in the ADS—i.e., the willingness of the driver (or passenger) to be vulnerable to the actions of the ADS—is essential if the driver is to leverage the opportunity to accomplish any NDRT [92]. Drivers must trust the ADS to feel comfortable dis-

engaging from the driving and focusing on the NDRT. Drivers who do not trust the ADS enough are less likely to hand over the driving to the ADS and fully disengage from the driving, shifting their attention to the NDRT. The lack of trust limits one's ability to fully engage in the NDRT. Therefore, it comes as no surprise that there has been extensive research on promoting drivers' trust in ADSs [3, 12, 78].

Advances have been made in understanding both the promotion of ADS trust and its impact on NDRT performance, but the influence of risk on that impact remains understudied. This is especially problematic as researchers readily admit that the use of ADSs is often accompanied by some level of risk [63, 97, 110]. Risk is defined as the degree of uncertainty associated with a given outcome [100], and is an important factor in trust-related phenomena because it has been found to determine whether or not trust translates into actual trusting behaviors [19, 46, 70]. Surprisingly, not much work has been directed at understanding the role of risk in ADS trust development and its impacts on drivers' trusting behaviors. Motivated by this lack of knowledge, the first goal of this dissertation is to provide analyses on the influence of internal and external risk factors on ADS trust and corresponding trusting behaviors. These analyses are presented in Chapter III, which is based on [10].

1.1.2 How to Estimate Drivers' Trust in AVs?

Trust is a highly abstract concept, and this abstractness makes measuring (or estimating) trust a challenging task [67]. Popular measures of trust are typically self-reported Likert scales based on subjective ratings. For example, individuals are usually asked to rate their degree of trust on a scale ranging from 1 to 7 [16, 43, 81]. Although self-reports are a direct way to estimate trust, they have several limitations. First, self-reporting is affected by peoples' individual biases, which makes a precise trust quantification hard to achieve [95]. Second, it is difficult to obtain repeated and updated estimates of trust over time without stopping or at least interrupting the

task or activity someone is engaged in [23, 136]. Specifically, it is not reasonable to expect ADSs to repeatedly interrupt drivers and ask them to complete a trust survey. As such, self-reported trust estimates are not an approach that can be relied on to assess drivers' trust in real-time.

In Chapter IV, this dissertation proposes an alternative approach to estimating drivers' trust by observing drivers' real-time actions and behaviors. The proposed method overcomes the limitations of previously published trust estimation approaches. For instance, those approaches fail to provide trust estimates in scales traditionally used for trust in automation [1], or require prohibitive sophisticated sensing and perception methods [1, 72]. Those approaches are also considered overly complex, as they include the processing of psychophysiological signals (e.g., galvanic skin response) that are not practical for the vehicular environments where driver-ADS interactions take place. Considering the potential implications for ADS and the far-reaching importance of trust to HRI research, the lack of robust methods for trust estimation is a significant gap to be filled. Especially in the case of self-driving vehicles, the ability to indirectly estimate trust opens several design possibilities, particularly for adaptive ADSs capable of conforming to drivers' trust levels and modifying their own behaviors accordingly. For example, trust estimations could be used in solutions for issues related to trust miscalibration—i.e., when drivers' trust in the ADS is not aligned with system's actual capabilities or reliability levels [21, 60, 81]. This possibility leads to the next high-level question to be addressed in this dissertation.

1.1.3 How to Influence and Calibrate Drivers' Trust in AVs?

In the future, automated systems will be expected to become aware of humans' trusting behaviors and to adapt their own behaviors, seeking to improve their interaction with humans [124]. One way to implement those adaptive capabilities is to develop methods for trust *management*, which we consider to be a robot's ability to

estimate and, if needed, to re-calibrate a human's trust in that robot. Trust calibration has recently become an important topic in human-robot interaction [21]. Recent calls have been made to better understand the problems associated with *overtrusting* and *undertrusting* automation and robots [107, 117]. In particular, the use of ADSs has been consistently plagued by problems associated with overtrusting and undertrusting automated capabilities.

Trust miscalibration is defined as a mismatch between a human's trust in an automated system and the capabilities of that system [59, 81]. Trust miscalibration is characterized by *overtrusting* or *undertrusting* an automated system, and it can harm the performances associated with the use of that system. Overtrusting an automated system can lead to *misuse*, where the human user relies on the system to handle tasks that exceed its capabilities. Undertrusting an automated system can lead to *disuse*, where the human fails to leverage the system's capabilities fully. Proper trust management can avoid both misuse and disuse of the automated system by estimating and, if needed, influencing the human's trust in the system to avoid trust miscalibration.

The ability to manage trust and avoid miscalibration is especially crucial for automated systems that can put people's lives at risk, such as AVs. Either misuse or disuse of an AV is a risk to the performance and safety of the team formed by the driver and the AV. Considering the current technology race in the automotive industry for AV development [125], AVs that can manage drivers' trust are a significant—if not urgent—demand.

In the driver–AV interaction context, the goal of trust calibration is to align the driver's trust to appropriate levels through a trust influence mechanism, for instance, by adapting the communication between the AV and the driver. The challenge of designing a trust calibrator, however, has not received as much attention from researchers. This research gap motivates the implementation of a trust management

system based on a trust calibrator that is presented in detail in Chapter V.

1.1.4 How to use trust to assign tasks between a human and an automated system?

Trust pervades people's relationships with other people, with organizations, and with machines [11, 60, 69, 82]. Trust relationships usually involve two types of agents: the trustor (the one who trusts) and the trustee (the one to be trusted). Trust depends on both the trustor's and the trustee's characteristics and is revealed when the trustor takes the risk of being vulnerable to the trustee's actions [69].

Researchers from the HRI field have proposed predictive trust models that try to capture how a human trustor develops trust in a robotic trustee [115, 132, 137]. A perspective that is generally overlooked, however, is how trust from a robotic trustor should develop over interactions with a trustee agent. This dissertation distinguishes between human trust, also called *natural trust*, from robotic trust, also called *artificial trust*. Therefore, current trust models are focused on natural trust and are useful for trust-aware decision-making, which requires the robot to estimate the human's trust in the robot to plan actions in a HRI setting. For example, trust-based partially observable Markov decision processes (POMDP) have been used by robots to plan actions while processing their human teammate's trust in applications involving robotic manipulation [17] and automated vehicles (AV) route planning [111].

Nonetheless, existing trust models have several shortcomings that hinder their ability to predict humans' natural trust and limit their application for robots' artificial trust computation. First, current trust models are limited in their ability to characterize the tasks that should be executed by trustees. Tasks must be characterized in terms of what capabilities and which proficiency levels (in those capabilities) are required from trustees to execute those tasks. For instance, driving requires certain levels of cognitive, sensory and physical capabilities from drivers [2]. Second, current

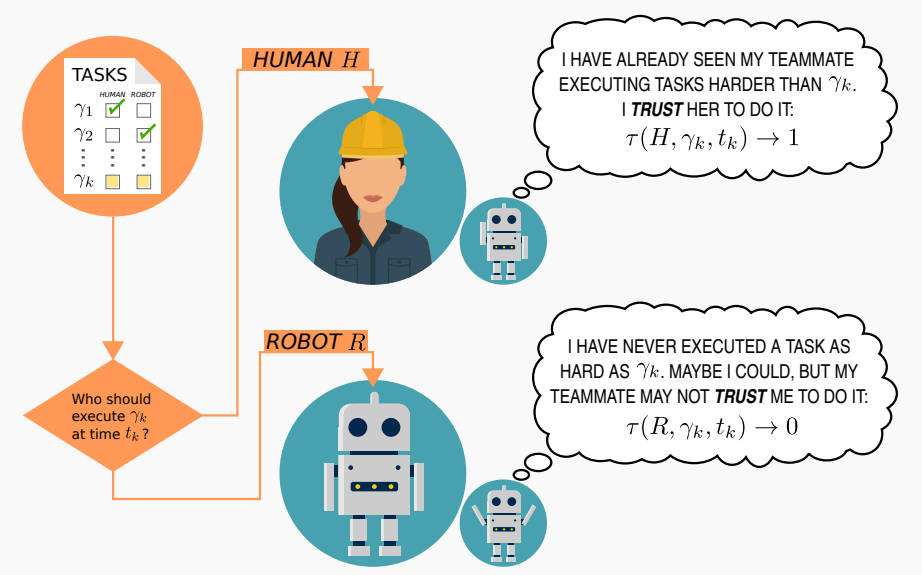

Figure 1.2: A team formed by human H and a robot R that collaborate executing tasks sequentially. Each task must be executed by one of the agents. The joint decision on which agent should execute each task depends on comparisons between the human's trust in the robot and the robot's trust in the human. A bi-directional trust model can be used for predicting a human's trust in a robot to execute a task, and to predict how much humans can be trusted to execute a task.

trust models also fall short of describing the trustee agents in terms of their proven capabilities. Trustees' capabilities characterization and quantification are important because,when a trustor knows that the trustee can (or can not) supply the types and levels of capabilities that a task demands, the trustor's trust in the trustee to execute that task is higher (or lower). Finally, because of a lack of trustee capabilities characterizations, current trust models are applicable for natural trust, or understanding human trust in a robot, but not for artificial trust, or determining how a robot should trust a human. Existing models are performance-centric and ignore "non-performance" trustees' factors, which are needed for artificial trust. To accommodate both natural and artificial trust in (human or robotic) trustees, a computational trust model must consider assessments of a trustee's non-performance capabilities, such as benevolence or integrity levels [69]. Therefore, although being sufficient for planning algorithms, existing trust models can not be used in more sophisticated control au-

thority allocation applications, which are likely to be based on comparisons between the human's trust in the robot and the robot's trust in the human [8], such as it would be needed in the situation represented in Figure 1.2. With this motivation, Chapter VI focuses on the development of a bi-directional trust model that can be used both for estimating a human's natural trust in a robotic system and a robotic system's (such as an ADS) artificial trust in a human (such as a vehicle driver).

1.2 Contributions

This section describes the main contributions of this dissertation.

1. *Investigation and characterization of how risk factors affect drivers' trust in ADSs*

   The study presented in Chapter III had two goals: first, to examine the impact of two types of risk—namely, internal and external risk—on ADS trust; second, to examine whether either type of risk weakens the impact of ADS trust on trusting behaviors such as ADS monitoring and NDRT performance.

   Results of that study showed that internal risk (low reliability ADS) reduces ADS trust but that external risk (low visibility) does not. In addition, internal risk moderated the positive impact that ADS trust had on NDRT performance. The positive impact of trust on NDRT performance was more prominent when the ADS was reliable (low internal risk). Moreover, external risk was found to moderate the impact of ADS trust on driver monitoring. ADS trust decreased monitoring when visibility was high (low risk) but not when visibility was low (high risk).

   The first contribution of this dissertation is the characterization of the role of risk in understanding the impacts of ADS trust on drivers' trusting behaviors. That contribution is represented by the conclusions of the study presented in

Chapter III, which can be summarized as follows. First, the specific type of risk (i.e., internal vs. external) matters. Only internal risk had a significant negative impact on ADS trust, while external risk had no significant impact on ADS trust. Therefore, future studies on considering risk and ADS trust should be careful to articulate the particular type of risk they are examining. Second, the moderating effects of internal and external risks on the impact of ADS trust on trusting behaviors such as driver monitoring and NDRT performance are demonstrated. High internal risk, which was represented by a faulty ADS, diminished the expected increase in performance when ADS trust is high, which indicates the occurrence of overtrust in driver-ADS interactions. External risk, which was represented by severely foggy weather, prevented drivers from reducing their ADS monitoring, even when those drivers reported high ADS trust.

2. *Development of a method for real-time drivers' ADS trust estimation*

   As mentioned in Subsection 1.1.2, keeping track of drivers' trust in ADSs in real-time without directly asking drivers to report their trust levels is crucial but challenging. To address this gap, this dissertation proposes a framework for estimating drivers' trust in ADSs in real-time, presented in Chapter IV.

   The proposed trust estimator is **the second contribution** of this dissertation, and is based on observable measures of drivers' behaviors and trust dynamic models. Although different trust estimation approaches have been previously reported in the literature [1,72], the proposed trust estimation method is simpler to implement. The proposed trust estimator represents trust in a continuous numerical scale, which is consistent with Muir's scale [83] and, therefore, also consistent with the theoretical background on trust in automation. Moreover, the proposed estimation framework relies on a discrete, linear time-invariant

(LTI) state-space dynamic model and on a Kalman filter-based estimation algorithm. This formulation makes the proposed trust estimation framework appropriate for treating the unpredictability that characterizes drivers' behaviors and for the design of innovative trust controllers. The trust estimator is intended to provide a means for the self-driving vehicle's ADS to track drivers' trust levels over time. It enables tracking drivers' trust levels without the need for directly requesting drivers to provide self-reports, which can be disruptive and impractical [67].

3. *Development of a method for drivers' ADS trust calibration*

   Chapter V presents the design of a trust calibration method, which is combined with the trust estimator to structure a framework for managing trust in AVs. The proposed trust calibrator is **the third contribution** of this dissertation, and focuses on how to re-calibrate drivers' trust after a trust miscalibration has been identified. The trust calibrator identifies trust miscalibration issues from the comparison of trust estimates with the capabilities of the AV and adjusts how the AV communicates with the driver. It compares the AV's capabilities with the driver's trust estimates to identify trust miscalibrations, and modifies the interactive behavior of the AV accordingly. The AV is the element that directly interacts with the drivers, providing verbal messages intended to influence drivers' trust in the AV, as illustrated in Figure 1.3. The trust calibrator is validated with a user study that shows that the proposed management framework was successful in its intent, being able to increase trust levels when drivers undertrusted the system and to decrease trust levels when drivers overtrusted the system.

4. *Development of a bi-directional trust model*

   Dynamic task allocation problems can benefit from the development and appli-

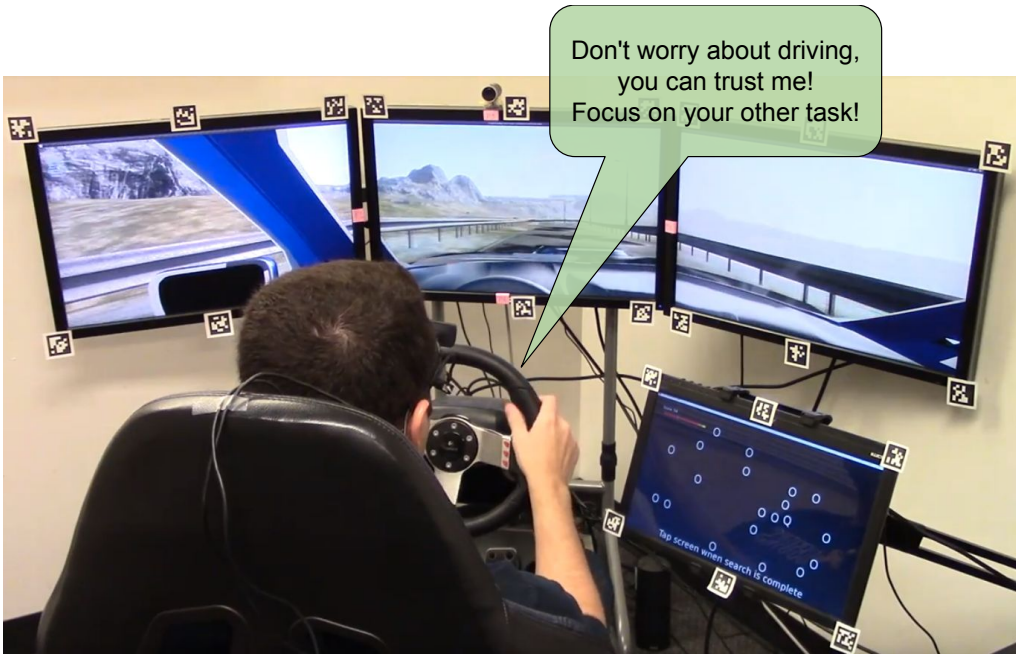

Figure 1.3: An undertrusting driver is encouraged by the AV system simulator to focus on his non-driving-related task (NDRT), to increase his trust level. An analogous situation would take place if the driver overtrusted the AV's capabilities, with the system then demanding his attention to the driving task.

cation of a bi-directional trust model able to accommodate both human's trust and robotic system's trust in (human or robotic) agents. This dissertation proposes a unified bi-directional trust model that characterizes tasks to be executed by potential trustee agents on a set of standard capability requirements. Then, trustee agents' capability profiles are built based on those trustee's performance on tasks they have executed previously. Trust is represented by the *probability that an agent can successfully execute a task, considering that agent's capability profile* (built after observations). By considering both the agent's capabilities and the task requirements, the proposed bi-directional model is applicable for determining a robot's artificial trust in a trustee agent. Moreover, the model can also be used for predicting trust transfer between tasks, similar to the model proposed in [115]. Chapter VI presents the development of this model, which was validated in a human subjects online experiment which resulted in a dataset

relating trust and task capabilities measurements. Therefore, **the fourth contribution** of this dissertation is the development of a new trust model that (i) can be used for the *artificial* trust computation and (ii) outperforms existing models for multi-task *natural* trust transfer prediction.

1.3 Dissertation Overview

This dissertation explores how trust can be used in the solutions of problems that are relevant in the human-robot interaction context (or, more specifically, in the driver-ADS interaction context). Chapter II presents an overview of the literature on trust, with a focus on the details that are important for answering the four motivating questions presented in Section 1.1. Chapter III goes deeper into identifying factors that affect drivers' trust in ADSs (ADS trust), focusing on the impacts of risk on ADS trust and on the relationships between trust and trusting behaviors. Chapter IV presents a method for real-time trust estimation in vehicular environments that considers and processes the driver's behaviors that reflect ADS trust. Chapter V leverages the trust estimation method from the previous chapter and presents a new framework for trust management combining trust estimation and a method for trust calibration. Chapter VI describes a novel model for bi-directional trust that can capture both natural and artificial trust (i.e., trust from a human trustor and trust from a robotic trustor), and that can be applied not only for the driver-ADS interaction context but also for a more general human-robot interaction context. Finally, Chapter VII wraps up the contributions of this dissertation and recommends directions for future research.

CHAPTER II

Background

This chapter presents an overview of trust in human-robot interactions, with a special focus on the particular case where the robot is a vehicle equipped with an automated driving system and the human drives the vehicle when necessary. Four main directions, characterizing the theoretical background in the key research topics of this dissertation, are discussed. Section 2.1 presents the literature relative to drivers' trust in automated driving systems, and its relation with drivers' risk perception and drivers' behaviors that reflect their trust (trusting behaviors). Section 2.2 focuses on trust estimation and Section 2.3 focuses on trust calibration, presenting relevant works that investigate how to measure and manage trust in real-time applications. Section 2.4 describes the recent research advances on trust computational models that are used in human-robot interaction applications and allow robots to reason about their human counterpart's trust levels and use trust to improve their collaboration.

2.1 Risk and Trust in Automated Driving Systems

2.1.1 ADS Trust and Trusting Behaviors

Trust has been conceptualized and utilized across different domains of research. Examples include user interface design for automotive applications [76, 85]; human factors and ergonomics [54, 83, 104]; and human–robot interaction [16, 34]. In this

dissertation, we consider ADS trust to be the willingness of the driver to be vulnerable to the actions of the ADS. More specifically, ADS actions represent the system's ability to drive the vehicle and to alert the driver about hazards that require the driver to take control. This "willingness to be vulnerable" is based on the drivers' attitude that the ADS in question will help them achieve their goals [54,92]. Trust is history-dependent and contingent upon drivers' prior knowledge about the capabilities and limitations of the ADS [49]. Reliance, differently from trust, occurs when drivers willingly cede control to the ADS [63]. ADS trust is vital for understanding when drivers will or will not rely on the ADS. A recent study [57] investigated ADS trust and reliance with six participants riding in a real-world self-driving vehicle. That study found that participants failed to fully trust the ADS even after 6 days of riding. In this regard, the ceding of control, as well as the degree of disengagement from the driving, can both be considered trusting behaviors [29, 121, 128].

Too much ADS trust is also a situation to be avoided. Overtrust occurs when the driver's ADS trust exceeds the ADS's capabilities. Trust is important because it influences drivers' behaviors directly, affecting their propensity to monitor the system and their ability to execute an NDRT [52]. Overtrust, however, leads to a higher chance that automation errors will go unnoticed and result in more accidents [79, 89]. To avoid this, drivers need to calibrate their ADS trust, aligning it with the system's capability [49, 87].

ADSs allow drivers to disengage from driving and engage in NDRTs safely. In the absence of ADSs, NDRTs are viewed as distractions that can lead to accidents [28]. However, the ability to engage in NDRTs by allowing the ADS to drive is increasingly viewed as a benefit [92,94,113]. As a result, researchers have been exploring the factors that promote better NDRT performance [53, 92]. For example, one such study [53] focused on selecting the most effective vehicle interface to support NDRTs.

Driver performance on NDRTs can be considered a trusting behavior induced

by ADS trust. However, the NDRTs must be carefully designed and meet specific requirements for NDRT performance to reflect trust. In general, the NDRT can not be as easy as to permit a high frequency switching between the NDRT the driving task—it must reach the driver's attention resources capacity. Additionally, the NDRT must be structurally similar to the driving task to increase multi-tasking difficulty. As suggested by multiple resources theory [127], tasks do not necessarily compete for a single pool of demand-sensitive resources. Therefore, if the NDRT has structural differences as compared to driving, and loads different attentional resource modalities (e.g., auditory instead of visual), the driver's ability to multi-task is higher. In that case, time-sharing can become more efficient, and the driver can achieve high NDRT performance without the intrinsic necessity to trust the AV because s/he will be able to execute both tasks at the same time easily. Therefore, the visual search NDRT is appropriate for ADS trust studies because it forces the sharing of the driver's visual attention, which is the primary resource required from the driver for safely operating the vehicle.

Other trusting behaviors can be observed and measured with the use of eye-tracking technology. Eye-movement recordings indicate where a person's attention is being directed, with the fixation durations indicating the amount of processing at the point of regard. Fixations indicate moments when the eyes are relatively stationary, taking in or "encoding" information. In an encoding task, higher fixation frequency indicates greater interest in a particular area of interest (AOI). Other eye-tracking metrics can also indicate different variables, such as: (i) saccades, which indicate quick eye movements occurring between fixations and measure processing difficulty during encoding; (ii) scanpaths, defined by complete saccade-fixation-saccade sequences, which indicates the efficiency of visual search; or (iii) blink rate; and (iv) pupil size, both indicating cognitive workload levels. In particular, gaze measurements, defined as the sum of fixation durations within an AOI, are helpful to compare attention

between target regions. In our driver-AV interaction context, drivers must split their visual attention resources between the NDRT and the driving task. For this reason, we use gaze measurements—defined in Chapters III, IV and V as *NDRT focus*—as an indicative of drivers' trust in the ADS.

Several studies have found that ADS trust increases NDRT performance [52, 94, 116]. The logic is simple: the more drivers trust the ADS, the more they focus on the NDRT; in turn, the better they perform on the NDRT [92]. Recently, Petersen *et al.* found that when drivers were provided with contextual information, increasing their situational awareness, ADS trust had a strong positive impact on NDRT performance [92]. In another example [40], Helldin *et al.* investigated the impact of uncertainty on trust and takeover speed. They found that drivers who were provided with a better understanding of the automation's abilities performed better on NDRTs. Similarly, Körber *et al.* found that participants with higher trust in automation spent more time on their NDRT and less time looking at the road—also confirming the trusting behaviors previously described [52]. In summary, the literature has demonstrated a strong and positive impact of ADS trust on NDRT performance.

2.1.2 ADS Trust and Risk

Scholars agree that risk is fundamental to understanding trust but most have focused on the direct relationship between risk and trust. Zhang *et al.* found a significant negative correlation between risk and trust [138]. They classify risk into two classes, namely safety risk and privacy risk. They defined safety risk as the possibility of accidents and physical harm from a system malfunction, while privacy risk originated from the possibility that travel or behavioral data could be transmitted to other parties, such as the government, vehicle developers, and insurance companies without notice, or even be used against the users or be hacked by others. Notably, they found that the negative correlation between risk and trust was significant only for

safety risk and not for privacy risk. A study conducted by Verbene *et al.* found that ADS trust also increased when risk was reduced [121]. Yet, other works have focused on understanding when risk reduced ADS trust [37, 63]. For example, Gremillion *et al.* found that when the ADS performed poorly, drivers' trust decreased and they relied less on the automation [37]. Conversely, when the ADS performed well, drivers' trust increased and drivers relied more on the ADS.

The classical integrative model of organizational trust by Mayer, Davis, and Schoorman also highlighted the potential moderating role of risk between trust and trusting behaviors [70]. The perceived risk associated with a given outcome determined whether trust led an individual to engage in trusting behaviors. In their trust model, the impact of trust on trusting behaviors was stronger when more risk was associated with an outcome. This was empirically verified in the context of virtual teams by Robert, Denis, and Hung [100]. They verified that higher risk involved in a given situation led to a stronger correlation between trust and trusting behavior. In the context of ADSs, Liu, Yang and Xu examined the relationship between risk and ADS trust [66]. Similar to other studies, they found that perceived risk had a negative relationship with trust. However, unlike other studies, they called attention to the complexity of the interactions between risk and trust. More specifically, they called for more research to better understand and model how risk and ADS trust interact with each other.

Although the research summarized here is valuable, more is needed, as pointed out in [66]. The literature on trust suggests that risk is vital to understanding trust impacts. This dissertation seeks to add to the literature by examining whether risk undermines the impacts of ADS trust. Without a better understanding of risk in the context of ADS trust, researchers and designers lack insight into an important mechanism needed to design ADSs. Chapter III focuses on the relationships between two types of risk (internal and external) on three important outcomes: ADS trust,

NDRT performance, and ADS monitoring.

2.2 Modeling and Estimating Trust in ADSs

2.2.1 Trust in Automation and Trust in Robots

Trust in automation has been discussed by researchers since it was first identified as a vital factor in supervisory control systems [112]. Formal definitions of trust in machines came from interpersonal trust theories [11,96] and were established by Muir in the late eighties [81]. Muir identified the need to avoid miscalibrations of trust in decision aids "so that [the user] neither underestimates nor overestimates its capabilities" [81]. Her work was then extended by Lee and Moray, who used an autoregressive moving average vector form (ARMAV) analysis to derive a transfer function for trust in a simulated semi-automatic pasteurization plant [58]. The inputs for this model were system performance (based on the plant's efficiency) and faults. They later focused on function allocation problems, and found that the difference between trust and self-confidence is crucial for users to define their allocation strategies [59].

The theoretical background on trust in automation has formed the basis for the development of more specific *trust in robots* measurement scales. Schaefer developed a scale that relies on the assessment of forty trust items, related to the human, the robot, and the environment where they operate [105]. Yagoda [134] created a measurement scale considering military applications and defining a list of HRI-related dimensions suggested by experts with extensive experience in the field. Charalambous *et al.* gathered qualitative trust-related questions focusing on the industrial human-robot collaboration (HRC) niche and developed a trust measurement scale for that specific context [15].

In this dissertation, and especially in Chapter IV and Chapter V, we consider the widely accepted definition of trust as "*the attitude that an agent will help achieve an*

individual's goals in a situation characterized by uncertainty and vulnerability" [60]. This definition aligns with Muir's standard questionnaire for trust self-reporting, which we used for trust quantification. Trust in automation is distinct from reliance on automation. Trust is an attitude that influences human's reliance behavior, characterized by engaging in automation usage. Trust miscalibrations are likely to induce inappropriate reliance, such as automation misuse or disuse [60].

2.2.2 Trust Dynamics and Estimation

Castelfranchi and Falcone [14] define the main aspects of trust dynamics as: how do the experiences of the *trustor agent* (both positive and negative experiences) influence trust changes; and how the instantaneous level of trust influences its subsequent change. These aspects are especially important when a human agent (in this case, the *trustor*) interacts with a machine (i.e., the *trustee*). As in a dynamic system, trust evolution is assumed to depend on the trust condition at a time instance and on the following inputs represented by the trustor's experiences with the trustee [58]. Several works have considered these basic assumptions and presented different approaches for trust dynamics modeling. The argument-based probabilistic trust (APT) model establishes the representation of trust as the probability of a reliable action, given the situation and system features [20]. In the reliance model, reliance is considered a behavior that is influenced by trust [60]. The three-layer hierarchical model describes trust as a result of dispositional, situational and learned factors involved in the human-automation interaction [43].

A relevant approach for modeling the dynamics of trust is that of Hu *et al.* [45], who developed a linear state-space model for the probability of trust responses within two possible choices: trust or distrust in a virtual obstacle detection system. In addition to developing trust-related dynamic models, researchers have used different psychophysiological signals to estimate trust. For instance, extending Hu's work [45],

Akash *et al.* [1] proposed schemes for controlling users' trust levels, applying electroencephalography and galvanic skin response measurements for trust estimation. However, psychophysiology-based methods suffer from at least two drawbacks. First and foremost, when using the reported psychophysiological methods, trust is not directly measured. Rather, the results of that method are conditional probabilities of achieving two states (trust or distrust), given prior signal patterns. Although this is a reasonable approach, previous research suggests that trust should be directly measured and represented in a continuous scale [15, 48, 83, 105]. Second, the sensor apparatus applied in psychophysiology-based methods is intrusive and can influence users' performance negatively, bringing practical implementation issues in applications such as self-driving vehicles.

The work presented in this dissertation, especially in Chapter IV, differs from previous research in two ways. First, a model that has trust as a continuous state variable is proposed. In this model, trust is defined in a numerical scale consistent with Muir's subjective scale [83]. Second, a simpler trust sensing method is presented. This method relies only on eye-tracking as a direct measure of drivers' behavior. Other variables that are used for sensing are intrinsic to the integration between ADS and the non-driving-related task (NDRT) executed by the driver.

2.2.3 System Malfunctions impact on Trust Dynamics

When not working properly, machines that are used to identify and diagnose hazardous situations—which might trigger human intervention—can present two distinct malfunction types: false alarms and misses [118]. On the one hand, false alarms occur when the system wrongfully diagnoses nonexistent hazards. On the other hand, when the system can not identify the existence of a hazard and no alarm is raised, a miss occurs. These different error types influence system users differently [4, 74, 75, 139], and also have distinct impacts on trust. The influence of false alarms and misses

on operators' behaviors was investigated by Dixon *et al.* [27], who has established a relationship with users compliance and reliance behaviors. After being exposed to false alarms, users reduced their compliance behavior, delaying their response to or even ignoring alerts from the system (the "cry wolf" effect). On the contrary, after misses, users allocated more attention to the task environment [26, 126, 129].

It is clear that false alarms and misses represent experiences that influence drivers' trust in ADSs. As systems that are designed to switch vehicle control with the driver in specific situations, ADSs rely on collision sensors that monitor the environment to make the decision to request drivers' intervention. Therefore, while other performance-related factors—such as the ADS's driving styles [13] or failures on different components of the ADS—could affect drivers' trust, we consider that those collision sensors were the most relevant and safety critical elements in SAE level 3 ADSs. In Chapter IV, we introduce system malfunctions only in the form of false alarms and misses on the simulated vehicle's collision warning system, while keeping other factors such as the vehicle's driving style and other failure types unchanged and generally acceptable: the vehicle followed the standard speed of the road, and no other type of system failure occurred.

2.3 Trust Calibration

Trust calibration is as important as trust estimation and plays a fundamental role for the management of trust in driver-ADS interaction (or human-robot, in general). People's trust in an automated system must be well calibrated, which means it has to be aligned with the system's capabilities. Miscalibrated trust is likely to lead to the inappropriate use of the system and accidents [7, 51, 55, 65]. However, the evolution of automated systems into autonomous robots with powerful sensing technologies has paved the way for new trust calibration strategies. Researchers have proposed strategies for autonomous robotic systems to try to perceive and process humans'

trust, and modify their own behaviors to influence humans' trust when necessary [7,17, 111]. Current trust-aware autonomous robotics systems are indicative that traditional concepts related to trust in automation are evolving and being reexamined by the HRI community. In Chapter V, the objective of trust calibration is to manipulate drivers' trust in the AV for aligning trust with the AV's capabilities (i.e., avoiding trust miscalibration). Several studies have identified factors that significantly impact trust in AVs, and, therefore, could be used for trust manipulation purposes. The most important of these factors are situation awareness and risk perception, which are influenced by the ability of the AV to interact with the driver. For instance, enhancing drivers' situation awareness facilitates increased trust in AVs [92, 93]. On the other hand, increasing drivers' perception of risk reduces their trust in AVs [4, 94, 139]. The trust management framework proposed in Chapter V takes advantage of these studies' results, and seeks to influence trust by varying situation awareness and risk perception through verbal communications from the AV to the driver.

2.4 Bi-Directional Trust

2.4.1 Utilitarian Trust Definition

Several trust definitions have been proposed, generally pointing to the trustor's willingness to be vulnerable to the trustee's actions [60, 69]. In this work, we consider Lee's trust definition [60]. However, Kok and Soh have recently proposed the following (adapted) definition for trust: "given a trustor agent A and a trustee agent B, A's trust in B is a multidimensional latent variable that mediates the relationship between events in the past and A's subsequent choice of relying on B in an uncertain environment" [51]. In Chapter VI, we adopt this utilitarian view of trust, which is aligned with a focus on the trustor-trustee pair [69] history of interactions, and is useful for the development of human-robot collaboration planning and control

methods.

2.4.2 Trust Computational Models

In robotics applications, the main goal behind the development of trust models is to give a robot the ability to estimate its human counterpart's trust (in that same robot). Trust models are usually applied to determine how much a human trusts a robot to perform a task (such as in Figure 1.2, when the robot R is chosen to execute a task). The robot uses this estimate of human trust to predict the human's behavior, such as intervening and taking over the task execution. For example, trust models are used in different trust-aware POMDP-based algorithms that have been proposed for robotic planning and decision-making [16,61,111]. Their objective is to eventually improve the robot's collaboration with the human, using human trust as a vital factor when planning the robot's actions.

Planning and decision-making frameworks usually rely on the use of probabilistic models for trust, such as those proposed in [33, 38, 132]. Xu and Dudek proposed an online probabilistic trust inference model for human-robot collaborations (OPTIMo) that uses a dynamic Bayesian network (DBN) combined with a linear Gaussian model, and recursively reduces the uncertainty around the human operator's trust. OPTIMo was tested in a human–unmanned aerial vehicle (UAV) collaboration setting [132] and, although some dynamic models had been proposed before [23, 58], OPTIMo was the first trust model capable of tracking human's trust in a robot with low latency and relatively high accuracy. The UAV, with OPTIMo, was able to track the human operator's trust by observing how much the human intervened in the UAV's operation.

There have been other Bayesian models proposed since OPTIMo. These models include personalized trust models that apply inference over a history of robot performances, such as [33] and [38]. Mahani *et al.* proposed a model for trust in a swarm of UAVs, establishing a baseline for human-multi-robot interaction trust prediction [33].

Guo and Yang have improved trust prediction accuracy (as compared to Lee's AR-MAV model [58] and OPTIMo [132]) by proposing a formulation that describes trust in terms of Beta probability distributions and aligns the inference processes with trust formation and evolution processes [38].

Although all previously mentioned approaches for trust modeling represented important advances in how we understand and describe humans' trust in robots, they suffer from a common limitation. Those models depend on the history of robots' performances on unique specific tasks, and are not applicable for trust transfer between two different tasks. The issue of multi-task trust transfer was recently approached by Soh *et al.* [115], who proposed Gaussian Processes and neural methods for predicting the transferred trust among different tasks that were described with NLP-based text embeddings. One goal of the bi-directional trust model proposed in Chapter VI is to deepen that discussion and improve prediction accuracy for multi-task trust transfer by: (i) describing tasks in terms of capability requirements, and (ii) describing potential trustee agents in terms of their proven capabilities that can be used to transfer trust to another task.

CHAPTER III

Trust in Automated Driving Systems, Risk and Driver Trusting Behaviors

3.1 Introduction

Automated driving systems (ADSs) allow vehicles to engage in self-driving under specific conditions. Along with the potential safety benefits, the increase in productivity through non-driving-related tasks (NDRTs) is often cited as a motivation behind the adoption of ADSs. Although advances have been made in understanding both the promotion of ADS trust and its impact on NDRT performance, the influence of risk remains largely understudied. This chapter presents a within-subjects experiment conducted to fill that gap. A total of 37 licensed drivers used a simulator where internal risk was manipulated by ADS reliability and external risk by visibility, producing a 2 (ADS reliability) $\times$ 2 (visibility) design. The results indicate that high reliability increases ADS trust and further enhances the positive impact of ADS trust on NDRT performance, while low visibility reduces the negative impact of ADS trust on driver monitoring. Results also suggest that trust increases over time if the system is reliable and that visibility did not have a significant impact on ADS trust. These findings are important for the design of intelligent ADSs that can respond to drivers' trusting behaviors.

This chapter is based on the work published in [10]. The remainder of the chapter is organized as follows. Section 3.2 presents a research framework, describing hypotheses about the relationship between risk, ADS trust and trusting behaviors. Section 3.3 describes the methodology applied for the experiment that was designed to validate the hypotheses. Section 3.4 presents the results obtained in the experiment. Section 3.5 discusses the main findings of the study and how these findings fit in the literature on ADS trust and risk. Section 3.6 highlights the main limitations of the study presented, while Section 3.7 summarizes its conclusions and contributions.

3.2 Study on ADS Trust and Risk

Based primarily on the relationship between risk and trust, several hypotheses were developed in the context of an ADS and a driver performing an NDRT. The ADS is designed to support NDRTs by providing the driver with semi-autonomous driving capability and recommendations based on the current driving situation. The system is considered a Level 3 ADS, in accordance with the classification defined in the SAE J3016 standard [101], because: (i) the simulated vehicle can drive conditionally under specific situations, (ii) the driver is a fallback-ready user of the vehicle, receptive to ADS-issued requests to intervene, and able to take control and drive when necessary, and (iii) the system can issue a request for the driver to intervene. The ADS's recommendations are designed to help the driver know when she or he has to disengage from the NDRT and take over the driving from the ADS. Drivers also have the option to monitor the driving situation themselves and determine when they should take over the driving independent of the ADS's recommendations. We hypothesize about the implications associated with: (i) reducing the ADS's reliability by having it provide incorrect recommendations and (ii) reducing the visibility in the driving situation by providing foggy weather.

3.2.1 Risk and ADS Trust

Based on prior ADS literature [37, 66, 121, 138], it is hypothesized that increases in either internal or external risk (i.e., reduced reliability or visibility) should reduce ADS trust for several reasons. For internal risk, the reduced reliability should inherently decrease the level of trust someone has in the ADS, i.e., a less reliable ADS is a less capable ADS. In this case, less reliable means an ADS that provides incorrect recommendations on when the driver should take over the driving. Therefore, drivers who receive incorrect recommendations would be likely to view the ADS as less capable. This would reduce their expectations about the system's ability, hence reducing ADS trust. For external risk, reduced visibility increases the difficulty of the driving situation. In our case, we used foggy weather to reduce visibility, which might cast doubt on the ADS's ability to make correct recommendations (on when the driver should take over). As visibility decreases, drivers should be less likely to believe that the ADS can assess the situation and make correct recommendations. Taken together, increases in both internal and external risks in the form of a less reliable ADS and less visibility should decrease the driver's trust in the ADS.

Hypothesis 1: ***Low ADS reliability reduces ADS trust.***

Hypothesis 2: ***Low visibility reduces ADS trust.***

3.2.2 Risk, ADS Trust and NDRT Performance

Internal risk should moderate the impact of ADS trust on NDRT performance. Based on prior literature, when internal risk is low, we should expect increases in ADS trust to lead to better NDRT performance [52, 94, 116]. The more the drivers trust the ADS, the more they can engage in the NDRT and disengage from driving. A reliable ADS provides the driver with correct recommendations, helping the driver to make good decisions. This explains the positive link between ADS trust and NDRT performance [94]. However, when internal risk is high, we should expect increases

in ADS trust to have little impact on NDRT performance. Trusting an unreliable ADS can actually have negative consequences for the driver. An unreliable ADS provides incorrect recommendations, causing the driver to make poor decisions. As such, increases in ADS trust should be less likely to directly translate to better NDRT performance.

Hypothesis 3: ***ADS reliability moderates the impact of ADS trust on NDRT performance in the following ways:***

- *When ADS reliability is high, ADS trust* ***increases*** *NDRT performance.*
- *When ADS reliability is low, ADS trust has* ***little or no impact*** *on NDRT performance.*

External risk should also moderate the impact of ADS trust on NDRT performance. More specifically, low visibility should reduce the impact of ADS trust on NDRT performance. When visibility is low, drivers are likely to engage in monitoring irrespective of their trust in the ADS. Drivers attempt to double-check the driving situation even with the information provided by the ADS. Overall, this choice is likely to weaken the potential impact of ADS trust on NDRT performance. However, when visibility is high, drivers are more likely to rely on the ADS to sense the environment and drive safely. Therefore, when external risk is low, higher ADS trust should translate into better NDRT performance. When external risk becomes evident to the drivers, they do not achieve their best NDRT performance, even when they reportedly trust the ADS. In all, trusting an ADS when visibility is high is likely to have positive consequences for the driver, and less so when visibility is low.

Hypothesis 4: ***Low visibility due to foggy weather moderates the impact of ADS trust on NDRT performance in the following ways:***

- *When visibility is high, ADS trust* ***increases*** *NDRT performance.*

- *When visibility is low, ADS trust has **little or no impact** on NDRT performance.*

3.2.3 Risk, ADS Trust and Monitoring

Internal risk should moderate the impact of ADS trust on monitoring. Based on prior literature, when internal risk is low we should expect increases in ADS trust to decrease the driver's monitoring of the driving situation [41, 46, 52, 70]. The more drivers trust the ADS, the more likely they will be to focus on the NDRT and refrain from monitoring the driving themselves. However, when the ADS is unreliable, drivers are likely to engage in monitoring irrespective of their level of trust in the ADS. When this occurs, ADS trust should not reduce the degree of monitoring. Thus, increases in ADS trust should reduce monitoring when internal risk is low but not when internal risk is high.

Hypothesis 5: ***ADS reliability moderates the impact of ADS trust on monitoring in the following ways:***

- *When ADS reliability is high, ADS trust **decreases** monitoring.*
- *When ADS reliability is low, ADS trust has **little or no impact** on monitoring.*

External risk should also moderate the impact of ADS trust on monitoring. During driving conditions of high visibility, ADS trust should reduce monitoring. When visibility is high, drivers are more likely to trust and rely on the ADS than to engage in their own monitoring of the driving situation. This explains the negative impact of ADS trust on monitoring. However, similarly to Hypothesis 4, when visibility is low, drivers are more likely to monitor irrespective of their ADS trust. As stated previously, drivers will double-check the driving situation over and above the information provided to them by the ADS. Although this might not be a wise decision relative to NDRT performance, drivers are likely to monitor the driving situation regardless

of their reported trust in the ADS. Therefore, trust in the ADS would not decrease monitoring. In sum, trusting an ADS should be likely to reduce monitoring when visibility is high but not when visibility is low.

Hypothesis 6: ***Low visibility (due to foggy weather) moderates the impact of ADS trust on monitoring in the following ways:***

- *When visibility is high, ADS trust* ***decreases*** *monitoring.*
- *When visibility is low, ADS trust has* ***little or no impact*** *on monitoring.*

Figure 3.1 presents our research framework, indicating the impacts of one factor on the other and representing pictorially the hypotheses with the labels H1, H2, H3, H4, H5 and H6.

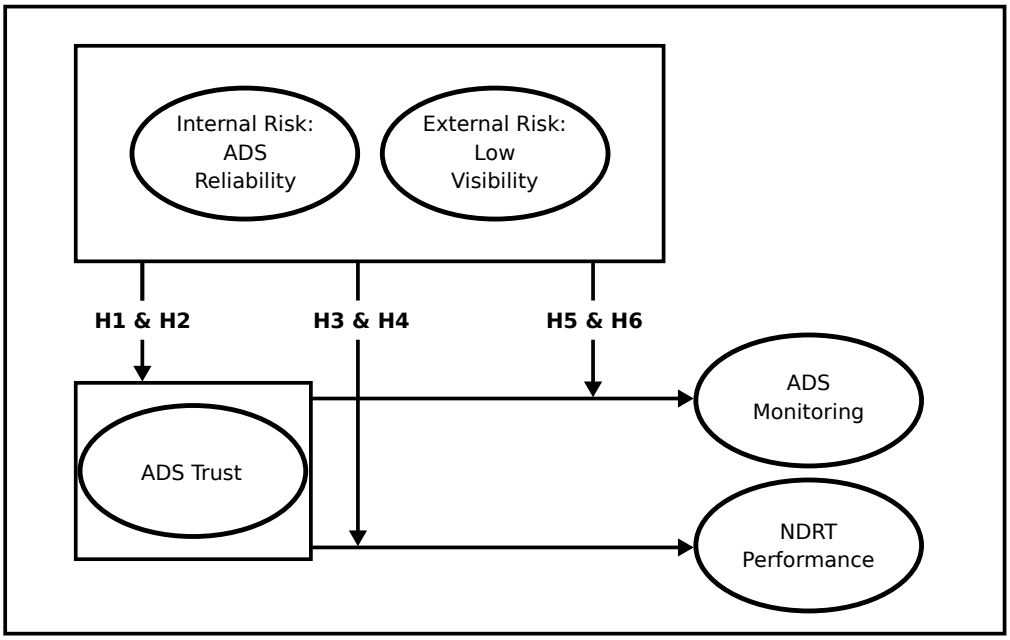

Figure 3.1: Research framework considered in this study. We hypothesized that **risks** reduce drivers' **trust** in the ADS. Moreover, **ADS trust** elicits **trusting behaviors** and promotes better **NDRT performance**. However, this relationship should be **influenced by the risks involved** in the context. ADS = automated driving system; NDRT = non-driving-related task.

3.3 Methodology

3.3.1 Participants

We recruited a total of 37 licensed drivers from the Ann Arbor, MI area to participate in the experiment. Participants were recruited via email advertising and printed posters. They were then directed to a website for eligibility screening. This screening required all participants to:

- be older than age 18,
- be a licensed driver,
- not be colorblind,
- have normal or corrected-to-normal vision (with contact lenses only–eye glasses were not allowed because they would interfere with the eye-tracker),
- have normal or corrected-to-normal auditory acuity,
- have no history of disorders or injuries that could affect their ability to use the simulator,
- not be military or civilian Department of Defense employees, and
- not have participated in the study before.

Participants' average age was 22.5 years (standard deviation [SD]=3.6 years), including 11 women, 25 men, and 1 participant who chose not to specify gender.

3.3.2 Experimental Tasks

3.3.2.1 Driving task

The primary task for the participants was to drive the simulated vehicle on the road with help from the ADS, while avoiding any collisions. The ADS provided

the following features to the driver: automatic lane-keeping, cruise control, forward collision alarm, and emergency braking. However, the vehicle was not able to switch lanes by itself. Participants could switch between AUTO mode (i.e., when the ADS was in charge of driving) and MANUAL mode (i.e., the participant was in charge of driving) at any point if they desired. The forward collision alarm was the only feature that did not work correctly in the unreliable ADS condition. The participants had to take active control to switch lanes and avoid hitting obstacle vehicles along the road. Figure 3.2 provides an example of the driving environment.

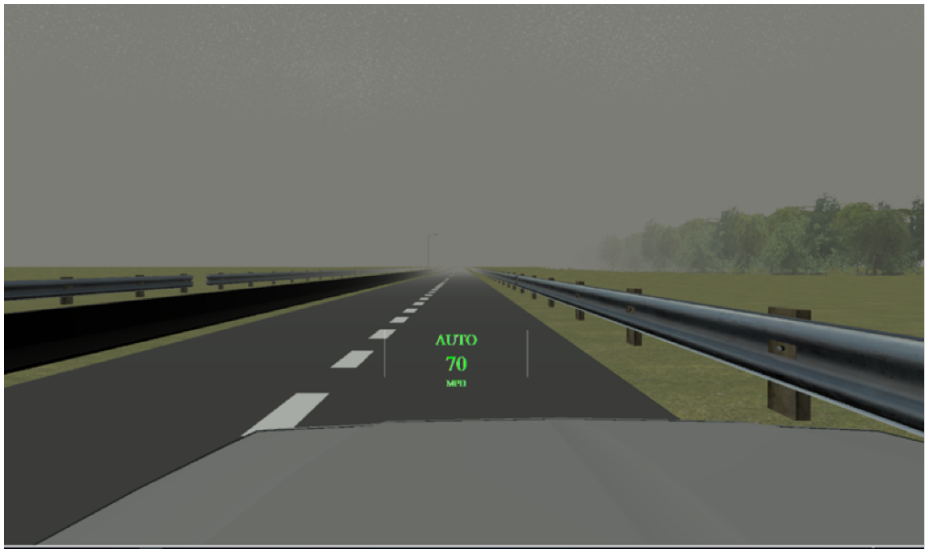

Figure 3.2: Driving task: to drive a vehicle on a highway and avoid the obstacles, with lane-keeping and alert assistance from the automated driving system.

Occasionally, the simulated vehicle alerted the participant that an upcoming parked vehicle was standing on the lane ahead. The alert system issued audible alarms. Alarms sounded two verbal messages: "stopped vehicle ahead,", displayed approximately 6.5 s before reaching a stopped vehicle, followed by "take control now," which sounded 5 s before reaching the obstacle. In those situations, if the participants did not take control in time, the emergency brake was triggered and prevented the collision. Participants received 10 alerts, representing 10 events per trial. In the unreliable ADS condition, these alerts were false alarms in three of the 10 events.

Figure 3.3 presents a typical order of events in a trial.

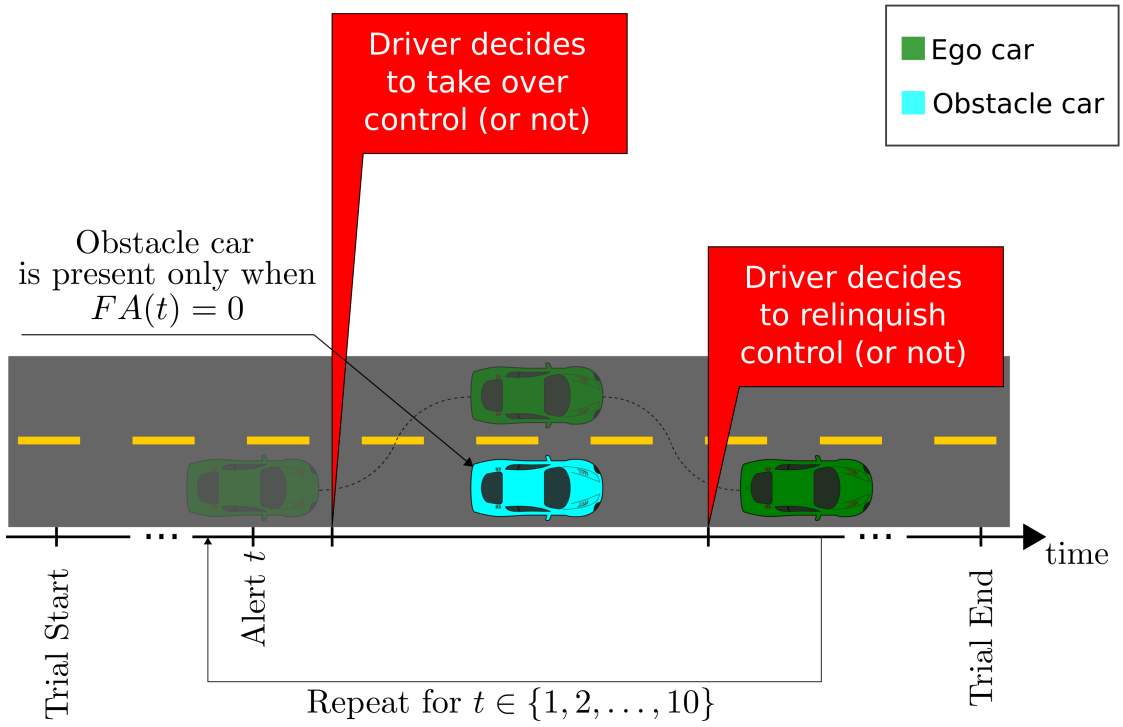

Figure 3.3: Timeline for one trial. Participants experienced all four trial conditions. Each trial had 10 alerts that could be true or false alarms. When the alert t was true, $FA(t) = 0$. When it was a false alarm, $FA(t) = 1$. Drivers were free to take over control at any time.

3.3.2.2 Non-driving-related task (NDRT)

The NDRT consisted of a modified version of the Psychology Experiment Building Language (PEBL) visual search task [119]. PEBL is a standard tool used by psychologists and social scientists to design and run behavioral tests [80]. In this task, participants used a touchscreen to repeatedly locate and select a target character (i.e., a "Q") that were placed among distractor characters (i.e., "O"s). Each time the participants correctly located and selected the target, they earned 1 point. Figure 3.4 provides a screenshot of the NDRT. As shown in Figure 3.5, the NDRT screen was positioned in a way to force the driver to choose between engaging in the

NRDT or monitoring the driving but not both. Additionally, each time the emergency stop was triggered to prevent a collision, drivers were penalized. The performance of the participants, represented by their final scores in the NDRT minus any penalties, was recorded for compensation purposes and to decide who was eligible to receive a monetary bonus. Participants received $15 and a cash bonus based on their performance. We promised a $5 bonus to the best performers under each risk condition, which encouraged participants to perform well in all four trials. Therefore, the NDRT functioned as a means of motivating participants to rely on the ADS. By doing so, participants were able to focus more on the NDRT and possibly receive the cash bonus. In addition, the loss of points from an emergency stop (and the consequent costs of losing cash bonuses) gave the participants a concrete sense of risk.

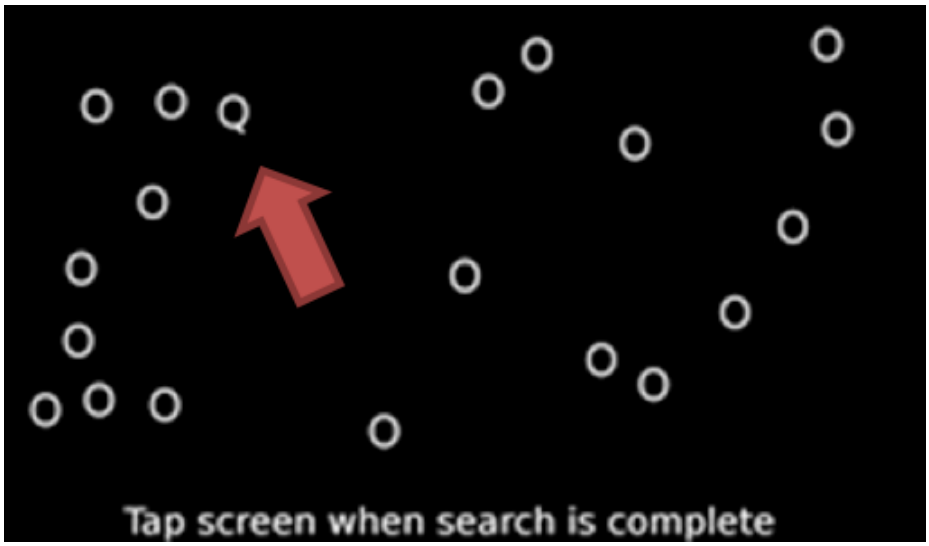

Figure 3.4: Non-driving-related task (NDRT): Visual search task where the participant had to find and point to the target "Q" among the "O"s. Each time participants correctly selected the target, they earned 1 point on their NDRT score. A penalty of 25 points was deducted from the NDRT score for each time the emergency stop was triggered. (The actual task did not show the red arrow.)

3.3.2.3 Apparatus

The simulator was composed of 3 LCD monitors integrated with a Logitech G-27 driving kit. A smaller touchscreen monitor was positioned at the right hand for the NDRT (see Figure 3.5).

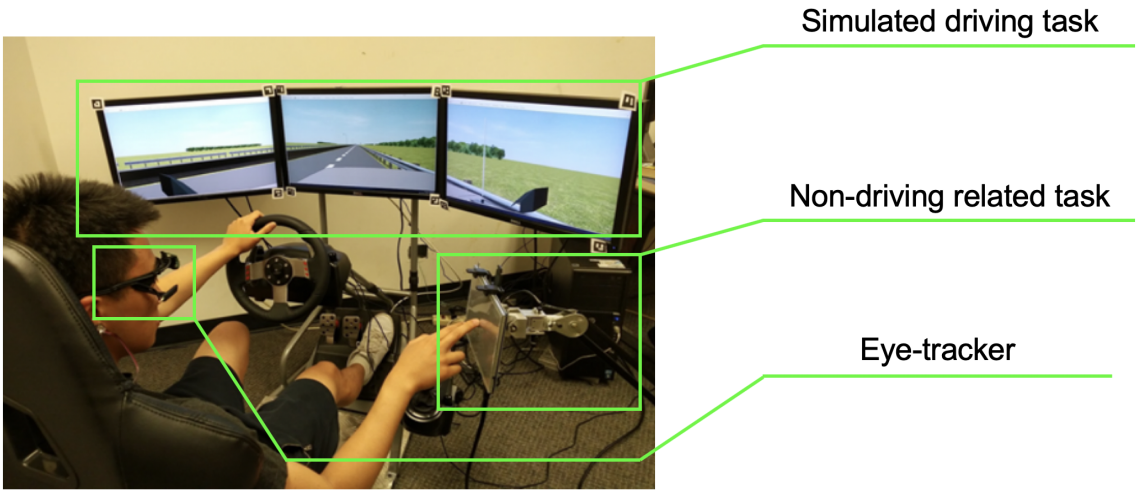

Figure 3.5: Experiment setup. The driving task was implemented with the Automated Navigation Virtual Environment Laboratory, or ANVEL [30]; the non-driving-related task (NDRT) was implemented with the Psychology Experiment Building Language, or PEBL [119]; Pupil Lab's Mobileye headset was the eye-tracker device used.

We developed the simulation with the Automated Navigation Virtual Environment Laboratory [30]. The console was placed to face the central monitoring screen so as to create a driving experience as close as possible to that of a real vehicle. For the eye-tracking device, we used Pupil Lab's Mobileye headset equipped with a fixed "world camera." This device acquired gaze positional data from participants' eyes as well as videos of the participants' fields of view and eye orientations.

3.3.3 Experimental Design

We employed a 2×2 within-subject design varying both the reliability of the automated driving system (ADS) and the visibility in the simulated environment. The ADS reliability was represented by two conditions: reliable (or perfect), when

the automation did not make any mistakes, and unreliable (or imperfect), when the automation gave some false alarms to the driver. The visibility was manipulated by two simulated weather conditions: clear or foggy. All conditions of the 2×2 design were experienced by all subjects.

ADS reliability and visibility were the two independent variables we manipulated to establish the 2×2 design. As stated, we manipulated the ADS reliability to assume two possible levels, represented by the reliable ADS $\times$ unreliable ADS conditions. We manipulated the reliability of the ADS by including false alarms. False alarms occurred when the ADS warned the driver of an obstacle on the road but, in fact, no obstacle was present. False alarms were the only system failures included in the simulation to manipulate the degree of ADS reliability. In the unreliable ADS conditions, false alarms occurred three times out of the ten alarms given to the driver per trial. In contrast, in the reliable ADS conditions, all ten alarms were correct. This percentage of false alarms (30%) is consistent with the prior literature [62, 94].

We also manipulated the simulated weather conditions to vary visibility in two levels. In clear weather, the high visibility permitted drivers to spot an obstacle $1,000$ ft (≈ 305 m) away, while the low visibility caused by foggy weather reduced this distance to 500 ft (≈ 152 m). The speed of the vehicle was regulated to 70 mph (≈ 113 km/h). Therefore, in terms of time to reach the obstacle, those distances represented time gaps of approximately 9.8 s in high visibility and 4.9 s in low visibility. The choice of visibility as a variable to represent the level of external risk involved in the driving context is consistent with prior literature. Low visibility levels have been found to increase the likelihood of rear-end collisions [135]. In addition, [56] found that users associated ADS risk with system errors or accidental events, rather than with psychological factors such as self-efficacy or ease of use, providing further support for both of this study's manipulations.

To introduce a notation that will be useful for the analyses of results, the binary

Boolean variables *Rel* and *Vis* were defined. These variables respectively represent the levels of ADS reliability and of visibility conditions in Equations (3.1) and (3.2).

$$Rel = \begin{cases} 0 & \text{if the ADS is 70\% reliable (unreliable ADS), and} \\ 1 & \text{if the ADS is 100\% reliable (reliable ADS).} \end{cases} \tag{3.1}$$

$$Vis = \begin{cases} 0 & \text{if the visibility is low (foggy weather), and} \\ 1 & \text{if the visibility is high (clear weather).} \end{cases} \tag{3.2}$$

In this study, *Rel* and *Vis* were static indicators in the sense that they did not vary during each trial. These variables represented the trial conditions and were set right before the start of each of the four trials experienced by the participants.

For the analysis of the evolution of some variables over the 10 alerts of each trial, a sequence $FA(t)$ was defined. $FA(t) = 0$ indicated that the ADS alarms worked properly at the alert t and, conversely, $FA(t) = 1$ indicated that a false alarm occurred at the alert t, $t \in \{1, 2, \ldots, 10\}$.

3.3.4 Measures

The following dependent variables were measured: (a) post-trial trust, (b) alert-wise dynamic trust, (c) risk perception variables, (d) final NDRT performance score, and (e) alert-wise dynamic monitoring ratio.

a) Post-trial trust, represented by T_{post}, was the numerical average of the answers to questions contained in the survey given to the participants after each trial (reproduced in Appendix A).

b) We also defined an alert-wise dynamic trust variable $T(t)$, which was computed with the increases or decreases in trust after each and every alert, including the false alarms (i.e., those for which $FA(t) = 1$). During the trial, subjects were asked after each ADS alert about their trust change, with the options of {decreased, no change,

increased}. The simulation was paused for some seconds while they answered the trust change question at the same tablet device they used for the NDRT. Their responses were translated to a quantized trust difference $\Delta T(t) \in \{-1, 0, 1\}$ respectively, for each event $t \in \{1, 2, \ldots, 10\}$.

To keep consistency between the post-trial trust and the dynamic trust, $T(t)$ is defined as in Equation (3.3),

$$T(t) = \begin{cases} T_{post} - \gamma \sum_{i=t+1}^{10} \Delta T(i), & \text{for } t \in \{0, 1, \ldots, 9\}, \text{ and} \\ T_{post}, & \text{for } t = 10. \end{cases} \tag{3.3}$$

Note that we defined $T(0)$ as the computed trust at the beginning of the trial, before any ADS alert. We chose the scaling factor $\gamma = 0.4$ to avoid negative values for the dynamic trust variable $T(t)$. To make sure that our findings would hold for different coefficients, we also computed the results for $\gamma = 0.2$, 0.3, and 0.5. All results involving the dynamic trust variable were consistent with the conclusions presented in section 3.4 for these γ coefficients.

c) Risk perceptions, represented by perceived reliability risk Rel_{perc} and perceived visibility risk Vis_{perc}, were also measured through standard surveys adapted from [100]. These can be found in Appendix A. These variables were used for a manipulation check, where we evaluated the participants' perception of how different were the risk conditions that they had experienced in each trial.

d) NDRT score (S_{NDRT}) was computed from each participant's total score obtained on the search task in each trial, where each correctly chosen "Q" was worth 1 point, and each emergency stop penalty deducted 25 points from the total.

e) Alert-wise dynamic monitoring ratio, represented by $r_m(t)$, was computed from the eye-tracking data to represent the eye movement properties [41]. When the participants switched their attention between the driving task and the NDRT, their gaze generally moved from the center monitor to the touchscreen and vice versa. Monitor-

ing ratio $r_m(t)$ was defined as the amount of time spent by the participant looking at the road (on the simulator monitors) during a time interval between the alerts $t-1$ and t, divided by this time interval.

All variables and their respective basic details are summarized in Table 3.1.

Table 3.1: Variable names and interpretations. Presented variables are extracted from experiment data and are used for linear mixed-effects models in the Results section.

Variable	Interpretation	Type	Set/Range
Rel	Reliability	Independent	{0, 1}
Vis	Visibility	Independent	{0, 1}
$FA(t)$	False alarm at alert t	Independent	{0, 1}
Rel_{perc}	Perceived reliability risk	Dependent	[1, 7]
Vis_{perc}	Perceived visibility risk	Dependent	[1, 7]
T_{post}	Post trial trust score	Dependent	[1, 7]
$T(t)$	Alert-wise dynamic trust score	Dependent	[0.2, 8.6] *
S_{NDRT}	Post-trial NDRT performance score	Dependent	{100, ..., 227} *
$r_m(t)$	Alert-wise dynamic monitoring ratio	Dependent	[0, 100%] *

Note: "*" denoted values observed from the data set. NDRT = non-driving-related-task.

3.3.5 Experimental Procedure

Upon arrival, participants signed a consent form to participate in the study. Next, participants completed a pre-experiment survey about demographics and their experience using driving assistance systems. This survey included questions about their risk tolerance and propensity to trust automated systems in general. Then, participants had a training session where they interacted with the simulator and performed the NDRT. The training drive allowed participants to become familiar with the simulator

and the NDRT prior to the four experimental conditions.

After the training session, participants were equipped with an eye-tracking headset, which was then calibrated. QR codes on each monitor allowed the eye-tracking software to determine which screen the participant was looking at. Next, the eye-tracking device was set up and participants started the first of the four trials. We counterbalanced the order of the trials to minimize any learning or ordering effects. For each trial, participants were tasked with both driving and performing the NDRT (described in subsection 3.3.2 Experimental Tasks). Participants were instructed to engage the automated driving mode as soon as they felt comfortable and start the NDRT, but not to totally neglect the driving (as the vehicle would ask them to take control). It took approximately 10 min for a participant to complete each trial. Finally, after each trial, participants completed a post-trial survey about their risk and trust perceptions. Participants were free to ask the experimenter for clarifications about the post-trial survey at any time. This survey used questions adapted from [83] (see Appendix A for the questions). After completing all four trials, participants were debriefed and received their compensation.

3.3.6 Analysis

We used linear mixed-effects (LME) models [131] to investigate the relationships among risk, trust, NDRT performance and monitoring ratios. The objective was to identify the parameters (represented by β) that significantly differed from 0 in each model. When β is significantly different from zero, we can consider that the associated factor influences the output variable. The errors associated with the models are represented by ϵ.

3.4 Results

3.4.1 Manipulation Check

We conducted a manipulation check for risk. We compared Rel_{perc} and Vis_{perc} between treatments with pairwise t-tests to determine whether the level of perceived risk differed significantly at the $\alpha = 0.001$ likelihood level. Table 3.2 shows that the means under each condition were significantly different from one another. Based on these results, we concluded that the manipulation was successful.

Table 3.2: Manipulation check for risk conditions.

Treatment Condition	Perceived Reliability/Visibility	Difference p-value
Low ADS Reliability ($Rel = 0$)	$Rel_{perc} = 2.10$	3.65×10^{-4} **
High ADS Reliability ($Rel = 1$)	$Rel_{perc} = 2.87$	
Low Visibility ($Vis = 0$)	$Vis_{perc} = 2.00$	1.40×10^{-9} **
High Visibility ($Vis = 1$)	$Vis_{perc} = 3.70$	

Note. ADS = automated driving system; NDRT = non-driving-related-task; Rel = reliability; Rel_{perc} = perceived reliability; Vis = visibility; Vis_{perc} = perceived visibility; Rel_{perc} and Vis_{perc} range: 1 to 7; ** $p < 0.01$.

3.4.2 Hypotheses Verification

The outcomes of the experiment were compared with our hypotheses, in order to validate them or not. The results are divided in three parts, directly linked to each

pair of hypotheses.

3.4.2.1 H1 and H2 – Impacts of risk on automated driving system (ADS) trust

To analyze the impacts of low reliability and low visibility on ADS trust, we built models considering both the post-trial trust T_{post} and the dynamic trust $T(t)$ as output variables.

For T_{post}, we fit the data to the model represented by Equation (3.4),

$$T_{post} = \beta_I + \beta_{Rel} Rel + \beta_{Vis} Vis + \epsilon \ , \tag{3.4}$$

where the obtained parameters and their respective significance values are presented in Table 3.3. As shown, ADS reliability significantly increased ADS trust, while visibility from the different weather conditions did not, thus supporting H1 but not H2.

Table 3.3: Parameters for the linear mixed-effects model of post-trial trust (T_{post}), with main effects for the independent variables *Rel* and *Vis*.

Factor affecting T_{post}, Equation (3.4)	Coefficient	S.E.	p-value
[Intercept]	$\beta_I = 4.88$	0.18	1.05×10^{-40} **
Reliability (*Rel*)	$\beta_{Rel} = 1.09$	0.14	1.60×10^{-11} **
Visibility (*Vis*)	$\beta_{Vis} = -0.06$	0.14	0.65

Note. S.E. = standard error; ** $p < 0.01$.

Similarly, for the dynamic trust $T(t)$, we built the model represented by Equation (3.5),

$$T(t) = \beta_I + \beta_{T(t-1)} T(t-1) + \beta_{Rel} Rel + \beta_{Vis} Vis + \epsilon \ , \tag{3.5}$$

to understand the influences caused by each risk type on the evolution of trust during a whole trial, considering the sequence of events indicated by t. In this model, however, we also considered the parameter $\beta_{T(t-1)}$, associated with the "one alert" delayed trust measurement $T(t-1)$. The parameters and their respective p-values are presented in Table 3.4.

Table 3.4: Parameters for the linear mixed-effects model of dynamic trust, or $T(t)$, with main effects for the delayed trust measure $T(t-1)$ and for the independent variables *Rel* and *Vis*.

Factor affecting $T(t)$, Equation (3.5)	Coefficient	S.E.	p-value
[Intercept]	$\beta_I = 0.274$	0.034	2.48×10^{-14} **
Dynamic (delayed) trust $T(t-1)$	$\beta_{T(t-1)} = 0.9597$	6.1×10^{-3}	1.46×10^{-39} **
Reliability (*Rel*)	$\beta_{Rel} = 0.083$	0.013	1.12×10^{-10} **
Visibility (*Vis*)	$\beta_{Vis} = -0.024$	0.012	0.036 *

Note. S.E. = standard error; * $p < 0.05$; ** $p < 0.01$.

The parameters from Table 3.4 show that ADS reliability has a significant effect on trust dynamics, and affects trust's evolution over time. Visibility's effect is also significant at the $\alpha = 0.05$ likelihood level. In summary, from the models represented by Equations (3.4) and (3.5) as well as their parameters, we observed that high ADS reliability had a significant positive impact on ADS trust. Visibility had a significant positive impact on Vis_{perc} and a significant negative impact on dynamic ADS trust, as shown in Table 3.4 and Equation (3.5). However, visibility did not have an impact on post trial ADS trust, as shown in Table 3.3 and Equation (3.4). Therefore, our first hypothesis was partially supported by our results.

These results are illustrated in Figures 3.6 and 3.7. Figure 3.6 presents the simulation of the model represented by Equation (3.5). For that simulation, we have

considered the initial condition $T(0) = 4$, which is the midpoint of the 7-point Likert scale. The use of a reliable ADS ($Rel = 1$) results in a faster increase in trust, while a low ADS reliability ($Rel = 0$) slows this evolution.

On the other hand, Figure 3.7 shows the average behavior for $T(t)$, considering the response data of all participants, for the different treatment conditions. The curves for which $Rel = 1$ follow the same pattern, indicating a solid trust increase over the usage time of a reliable ADS. Furthermore, the final values for $T(10)$, which corresponds to T_{post}, are not significantly different, both being close to 5.9 points. In low-reliability conditions ($Rel = 0$), the curves indicate decreases for specific alert indexes t, coincident with the false alarms provided by the ADS. That is, for $Rel = 0$ and $Vis = 1$, we had false alarms for $t = 3, 4, 6$ while for $Rel = 0$ and $Vis = 0$, false alarms occurred for $t = 2, 4, 5$. Moreover, for both low ADS reliability conditions, the average value of $T(10) = T_{post}$ was about 4.8.

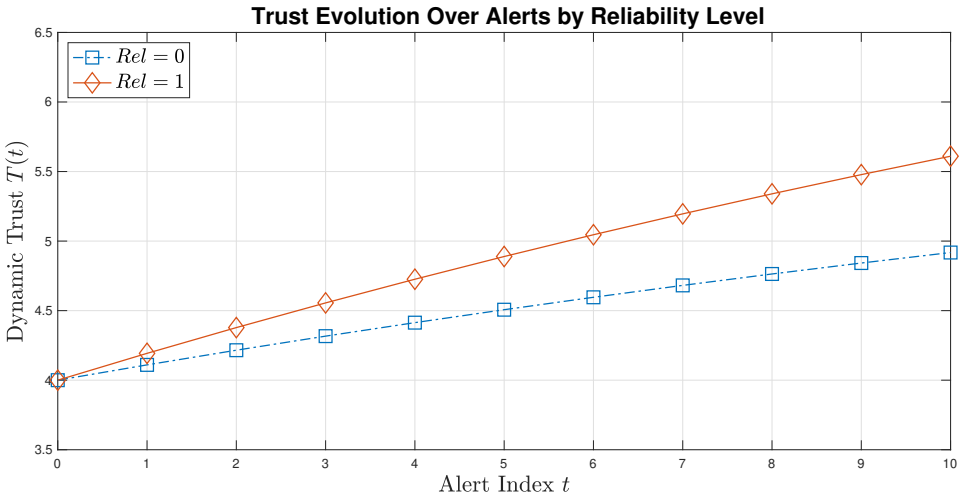

Figure 3.6: Curves illustrate the simulation of the model represented by Equation (3.5). We chose $T(0) = 4$ for both conditions to better compare the results. When $Rel = 1$ (i.e., when participants were using a reliable ADS), trust increased faster than when $Rel = 0$ (i.e., when participants were using an unreliable ADS). For both curves, $Vis = 0$.

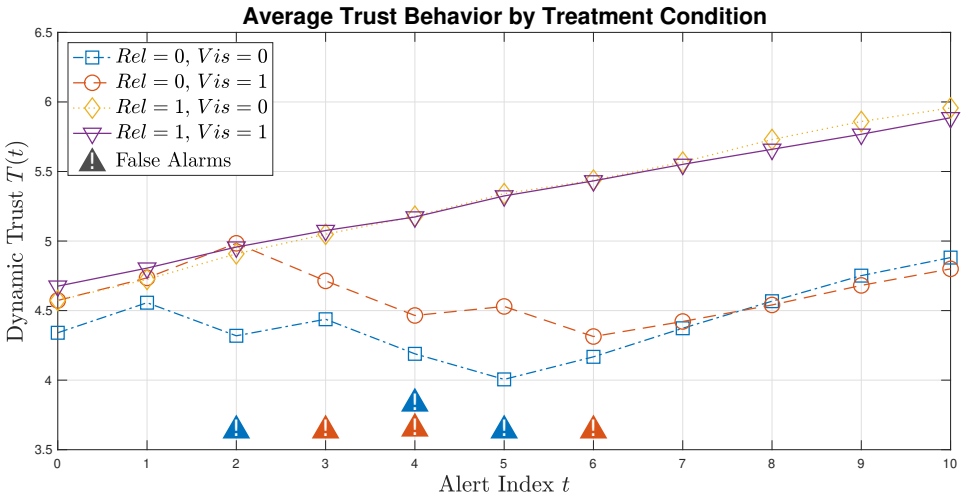

Figure 3.7: Plots of the average $T(t)$ for all participants for each reliability and visibility condition. When $Rel = 1$ (i.e., when participants were using a reliable ADS), $T(t)$ increased steadily over the alerts indicated by t. When $Rel = 0$ (i.e., when participants were using an unreliable ADS), the occurrence of false alarms resulted in decrements in $T(t)$. This happened for $t = 2, 4, 5$ when $Vis = 0$ and for $t = 3, 4, 6$ when $Vis = 1$. For these t, $FA(t) = 1$.

3.4.2.2 H3 and H4 – Influence of risk on the impacts of ADS trust on non-driving-related task (NDRT) performance

The second pair of hypotheses asserted that both low reliability and low visibility should moderate the impact of ADS trust on NDRT performance. This claim was only partially supported by our results, as we concluded by analyzing the model expressed in Equation (3.6) and its parameters listed in Table 3.5.

$$\begin{aligned} S_{NDRT} = \beta_I + \beta_{T_{post}} T_{post} + \beta_{Rel} Rel + \beta_{Vis}\, Vis + \beta_{T_{post}\times Rel}[T_{post} \times Rel] \\ + \beta_{T_{post}\times Vis}[T_{post} \times Vis] + \beta_{Rel\times Vis}[Rel \times Vis] + \epsilon. \end{aligned} \tag{3.6}$$

From the significant positive value for $\beta_{T_{post}\times Rel}$, we concluded that ADS reliability moderates the impact of ADS trust on NDRT performance (H3). The moderating effect of visibility represented by $\beta_{T_{post}\times Vis}$ was not significant (H4).

Table 3.5: Non-driving-related task score (S_{NDRT}) linear mixed-effects model parameters, with main effects for the post-trial average trust measure T_{post} and for the independent variables *Rel* and *Vis*, as well as their interaction effects. The interaction effects represent the moderating influence on the impacts of ADS trust on NDRT performance.

Factor affecting S_{NDRT}, Equation (3.6)	Coefficient	S.E.	p-value
[Intercept]	$\beta_I = 191$	14	9.44×10^{-25} **
Post-trial Trust T_{post}	$\beta_{T_{post}} = 3.1$	2.7	0.25
Reliability *Rel*	$\beta_{Rel} = -39$	19	0.045
Visibility *Vis*	$\beta_{Vis} = -4$	15	0.785
Interaction $T_{post} \times Rel$	$\beta_{T_{post} \times Rel} = 7.3$	3.2	0.028 *
Interaction $T_{post} \times Vis$	$\beta_{T_{post} \times Vis} = 1.7$	3.1	0.58
Interaction $Rel \times Vis$	$\beta_{Rel \times Vis} = -20.8$	7.6	0.008 **

Note. S_{NDRT} = non-driving-related task score; S.E. = standard error; * $p < 0.05$; ** $p < 0.01$.

Figure 3.8 represents the relationship corresponding to the results demonstrated by Equation (3.6) and its parameters (Table 3.5). With low reliability, the weaker slopes indicate that a higher ADS trust level did not result in a significantly better NDRT performance. When using a reliable ADS, however, the greater slope indicates that a higher trust corresponded to better performance.

3.4.2.3 H5 and H6 – Influence of risk on the impacts of ADS trust on monitoring ratio

H5 and H6 state that both low ADS reliability and low visibility should moderate the impact of ADS trust on monitoring ratio. These hypotheses are also partially supported by the model that relates $r_m(t)$ with the variables $T(t-1)$, *Rel* and *Vis*,

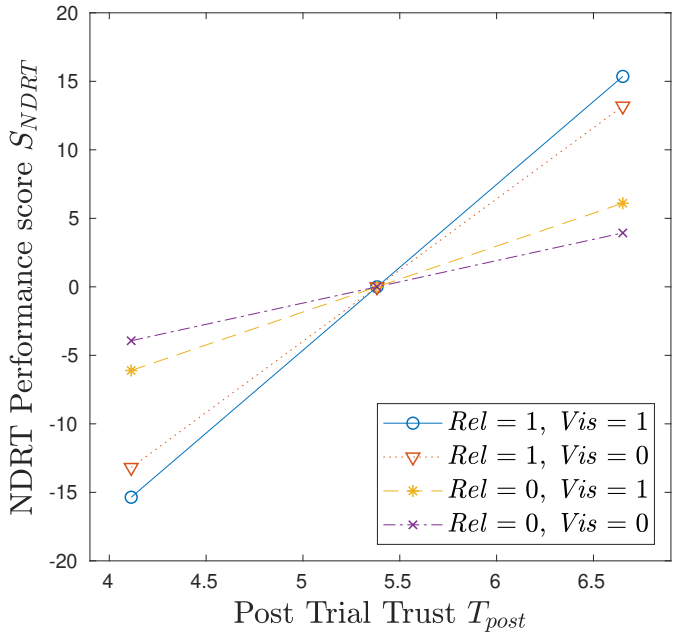

Figure 3.8: Correspondence between T_{post} and respective S_{NDRT} deviations around the mean. Here, the mean value for T_{post} is around $\mu = 5.4$, and the standard deviation is approximately $\sigma = 1.3$. The interval between one standard deviation above and below the mean ($\mu \pm \sigma$) is considered. The mean values for S_{NDRT} were all brought together at zero, for the comparison of slopes. For all conditions where $Rel = 1$, the slope is greater than when $Rel = 0$. Therefore, when using an unreliable ADS, participants could not translate a higher ADS trust level into significantly better NDRT performance. Visibility does not influence this relationship significantly. ADS = automated driving system; NDRT = non-driving-related task; Rel = reliability; Vis = visibility; S_{NDRT} = non-driving-related task score.

as we concluded from Equation (3.7) and its parameters (shown in Table 3.6). The use of $T(t-1)$ is justified because $r_m(t)$ was measured during the time period between alerts indexed by $t-1$ and t. Thus, we computed the impact of the trust responses on monitoring ratios measured right after the participants were asked about their trust changes.

$$r_m(t) = \beta_I + \beta_{T(t-1)}T(t-1) + \beta_{Rel}Rel + \beta_{Vis}Vis + \beta_{T(t-1)\times Rel}[T(t-1) \times Rel] + \beta_{T(t-1)\times Vis}[T(t-1) \times Vis] + \beta_{Rel\times Vis}[Rel \times Vis] + \epsilon. \tag{3.7}$$

Table 3.6: Monitoring ratio ($r_m(t)$) linear mixed-effects model parameters, with main effects for the delayed trust measure $T(t-1)$ and for the independent variables *Rel* and *Vis*, as well as their interaction effects. The interaction effects represent the moderating influence on the impacts of automated driving system trust on monitoring ratio.

Factor affecting $r_m(t)$, Equation (3.7)	Coefficient	S.E.	p-value
[Intercept]	$\beta_I = 0.403$	0.074	1.25×10^{-7} **
Dynamic (delayed) Trust $T(t-1)$	$\beta_{T(t-1)} = 0.006$	0.017	0.72
Reliability indicator *Rel*	$\beta_{Rel} = 0.013$	0.095	0.89
Visibility indicator *Vis*	$\beta_{Vis} = 0.144$	0.084	0.086
Interaction $T(t-1) \times Rel$	$\beta_{T(t-1)\times Rel} = -0.004$	0.018	0.83
Interaction $T(t-1) \times Vis$	$\beta_{T(t-1)\times Vis} = -0.041$	0.018	0.025 *
Interaction $Rel \times Vis$	$\beta_{Rel\times Vis} = 0.038$	0.048	0.42

Note. S.E. = standard error; * $p < 0.05$; ** $p < 0.01$.

The value of $\beta_I = 0.403$ in Table 3.6 indicates an average basic monitoring ratio for the participants, specifically when disregarding the impacts of trust and when $Rel = Vis = 0$. The results from Table 3.6 also show that monitoring ratio is negatively correlated with the interaction between $T(t-1)$ and *Vis*. That is, with high visibility (i.e., in clear weather conditions), the subjects trusted the ADS more, looked at the road less and focused on the secondary task more. However, under low visibility (i.e., foggy weather), such impact of trust was greatly reduced and monitoring ratio was no longer an effective trusting behavior. Reliability, however, had no significant impact on $r_m(t)$, nor did it moderate the impact of $T(t-1)$ on $r_m(t)$. These results corroborate H6 but not H5.

The relationship between $T(t-1)$ and $r_m(t)$ indicated by Equation (3.7) is illustrated in Figure 3.9, which summarizes all combinations of *Vis* and *Rel*. The figure shows that better visibility enabled a decrease in monitoring ratios when participants reported higher ADS trust. This is represented by the negative slopes when $Vis = 1$. Contrarily, when $Vis = 0$, this correlation became irrelevant, with the slope parameter assuming the value $\beta_{T(t-1)} = 0.006$, but with no significance.

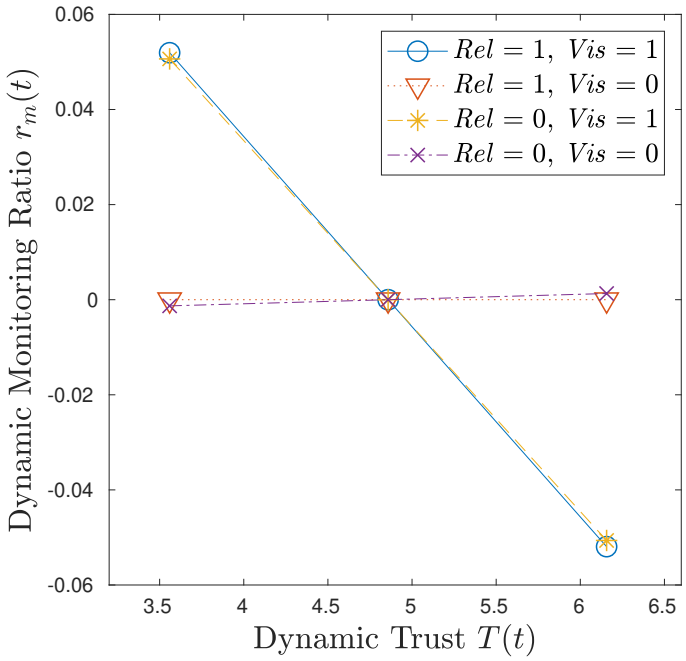

Figure 3.9: Correspondence between dynamic trust $T(t)$ and respective $r_m(t)$ deviations around the mean. Here, the mean value for $T(t)$ is around $\mu = 4.9$, and the standard deviation is approximately $\sigma = 1.3$. The interval between one standard deviation above and below the mean ($\mu \pm \sigma$) is considered, and the mean values for $r_m(t)$ were all brought together to zero, for the comparison of slopes. For all conditions where $Vis = 1$, the slope was negative, which did not happen when $Vis = 0$. The result shows that for $Vis = 1$, higher trust led to smaller monitoring ratios. In other words, high visibility allowed drivers to demonstrate their ADS trust by reducing system monitoring. However, when the visibility conditions were poor ($Vis = 0$), drivers did not decrease monitoring, even when they reported having higher ADS trust. ADS reliability did not influence this relationship significantly. *Rel* = reliability; *Vis* = visibility.

3.5 Discussion

The goals of this chapter were: (i) to investigate how different types of risk influence automated driving system (ADS) trust development, and (ii) to understand when different risk types undermine or strengthen the impact of automated driving system (ADS) trust on both non-driving-related task (NDRT) performance and monitoring ratio. Results of this study can be organized around three overarching findings. First, the use of an unreliable ADS reduced ADS trust (H1 supported), but foggy weather with low visibility did not (H2 not supported). This is consistent with what is shown in Figure 3.7, that on average trust increases over time if the system is reliable. Second, the use of an unreliable ADS moderated the positive impact of ADS trust on non-driving-related task (NDRT) performance (H3 supported), while low visibility did not (H4 not supported). Third, low visibility moderated the impact of ADS trust on monitoring (supporting H6), but low reliability did not (not supporting H5). Next, we discuss our contributions to the literature.

First, the findings here presented contribute to the cumulative research on the antecedents of ADS trust. Our first major finding is that the type of risk is important when understanding its effects on ADS trust. Research has suggested that, as risk increases, ADS trust decreases [37,121]. Since our manipulation check results confirm that our scenarios did induce higher perceptions of reliability and visibility (Table 3.2), our findings are consistent with prior literature for internal risk, represented by low reliability, but are not consistent with regards to external risk, represented by low visibility. Only low reliability resulted in lower ADS trust. Thus, our results extend the existing literature by demonstrating the distinct impacts of internal and external risks. Before [10], no studies had specifically distinguished between risk types and considered their influence on ADS trust.

Second, this study contributes to the literature by clarifying the boundary conditions on the impact of ADS trust on NDRT performance. A large body of research

has focused on the positive impacts of ADS trust on NDRT performance [52, 92, 116]. Our research extends prior work by showing when ADS trust is not likely to lead to better NDRT performance. Results of our study show that the positive impact of ADS trust on NDRT performance also depends on risk, and particularly on the type of risk. Our results are consistent with prior work when the ADS was working perfectly.

However, for an unreliable ADS, ADS trust had little or no impact on NDRT performance. External risk (represented by low visibility) did not significantly affect the relationship between trust and NDRT performance. Given our findings on the influence of risk in this relationship, we conclude that a highly reliable system is crucial for higher ADS trust to result in improved NDRT performance, whereas the visibility conditions in the environment are less important. These findings are novel because the existing literature has not explored the effects of risk from different sources on the impacts of ADS trust on NDRT performance.

Third, this study contributes to the literature by identifying the role of risk on the impact of ADS trust on monitoring. Specifically, this study found that the relationship between ADS trust and monitoring ratio also depends the type of risk. Prior research on ADS trust and monitoring has typically found that ADS trust reduces monitoring [41, 52]. When a driver trusts the ADS more, the driver spends less time watching the road. Our results were consistent with these established results only when there was high visibility in the environment. However, when the visibility was low because of severe fog, increases in ADS trust had almost no impact on monitoring. Whether ADS trust leads to less monitoring depends on the visibility levels; it does not depend on ADS reliability. Ironically, when drivers should be relying on the ADS the most (i.e., in low-visibility conditions), they apparently are not. These results were unexpected and provided a novel finding about the influences of risk on the relationship between ADS trust levels and monitoring. These results

also imply that an ADS that attempts to estimate the drivers' trust level based on the observed monitoring ratio cannot ignore the context presented by the external visibility conditions.

Finally, this chapter contributes to the ADS trust literature and has practical implications for the design of innovative ADS technologies. The relationships among trust, risk, NDRT performance and trusting behaviors could be incorporated in a trust estimation framework. As expected, our findings showed that unreliable ADSs (e.g., false alarms) could reduce driver trust in the system. An ADS that is self-aware when it has made a mistake might be able to explain to the driver what happened and, if not re-gain the driver's trust, at least help the driver to understand the limitations of the ADS. Intelligent ADSs could sense monitoring and performance and could benefit from our conclusions to estimate drivers' ADS trust more accurately. Our findings also indicate that monitoring ratio should be considered as a trusting behavior only when the environmental conditions permit—i.e., when weather is clear and visibility is high. Combining these trust estimates with sensed environmental conditions, intelligent systems can decide how to act to manage a driver's trust levels appropriately, attempting to avoid both over-trust and under-trust [37] which can both lead to serious problems.

3.6 Limitations and Future Research

The study presented in this chapter had several limitations. The first is related to our experimental setup: we used a simulated driving environment instead of a real vehicle. Participants could have different risk perceptions when an automated driving system (ADS) error could lead to a life-threatening accident instead of a monetary loss, and this could strengthen the relationships we found. Previous work has shown that individuals respond similarly to real and simulated environments [42], but the use of an actual vehicle in more realistic conditions could be the subject of future

research efforts.

We also manipulated our risk conditions varying only one internal and one external risk factor: ADS reliability and visibility according to weather conditions. ADS designers are expected to be very conservative regarding safety and, because of that, false alarms are more likely to be present in autonomous vehicles than misses. This is the reason why, although not being safety-critical, false alarms were chosen to represent flaws in system reliability in this work. However, to extend our conclusions, future research might specifically investigate the impact of different types of both internal and external risks. For internal risks, both false alarms and misses could be considered. For external risks, an extension of this work could be the introduction of rain or wet roads, not only reducing visibility but also affecting the ADS's and the driver's abilities to operate the vehicle. In addition, we only varied two levels of ADS reliability, 0 error or 30% error. However, future automated vehicles are expected to have much lower failure rates than 30%. Therefore, it would be important for future studies to consider examining the impact of lower error rates on ADS trust.

Another limitation is the demographic distribution of our participants. In our study, subjects were relatively young and most were men. Therefore, we should be cautious when expanding our conclusions to the general population. Additionally, personal traits have shown to impact user's trust in robots generally and automated vehicles specifically [86, 98, 99]. Future studies could examine how user's personality traits may influence ADS trust in the presence of risk.

This study did not employ explanations from the ADS to help the driver understand why the ADS did or did not work properly. Prior research had employed explanations as a means of promoting driver trust when unexpected events or actions took place. That being said, it is not clear that any research has examined the impacts of explanations relative to the effects of risk on trust. Future research could investigate the ability of explanations from the ADS to reduce uncertainty and risk.

In addition, such explanations can help drivers increase their ADS trust and predict when the ADS may or may not work properly [29, 31, 49]. Prior research has shown that drivers can still trust an unreliable ADS when they can predict when or why it might fail. Future studies should consider including the impacts of the driver's knowledge of the system to provide additional insights into the influence of risk on the impacts of ADS trust.

3.7 Conclusions and Contribution

In this chapter, we investigated how different risk types influence drivers' trust in automated driving systems (ADSs). We examined how risk moderates the impacts of ADS trust on drivers' trusting behaviors, and the impacts of ADS trust on their performance in a secondary, non-driving-related task (NDRT). The study here presented considered two risk types: internal, represented by low ADS reliability; and external, associated with low visibility from foggy weather. The three major findings were: (1) The negative impact of risk on ADS trust depends on the type of risk and, in particular, risks from external sources (such as foggy weather) did not have a significant negative impact on ADS trust. (2) The positive impact of ADS trust on NDRT performance depends not only on risk but also on the type of risk; for an unreliable ADS, ADS trust had little or no impact on NDRT performance. (3) The negative impact of ADS trust on monitoring ratio depends not only on risk, but also on the type of risk. When the visibility was low because of severe fog, ADS trust had almost no impact on monitoring ratio.

These findings characterize how risk factors affect drivers' trust in ADSs and, taken as a whole, represent the first main contribution of this dissertation. New ADS studies can take these findings into consideration to better understand how drivers' trust is related to their performance and behavior under different risk contexts. Risk influences the evolution of drivers' ADS trust and, ultimately, moderates their ability

to rely completely on the system and perform tasks other than driving. With new artificial intelligence and machine-learning-enabled technologies being able to identify and classify complex information and different contexts, the perception and processing of trust and risk are likely to become possible. Thus, a better understanding of how these factors evolve and influence each other is fundamental for the design of future intelligent ADSs.

Leveraging part of the knowledge obtained in this chapter, Chapter IV will present a trust estimation method that is based on a linear model relating internal risk (i.e., the occurrence of false alarms and misses in ADS technology) with ADS trust dynamics.

CHAPTER IV

Estimation of Drivers' Trust in ADSs

4.1 Introduction

Trust miscalibrations, represented by undertrust and overtrust, hinder the interaction between drivers and self-driving vehicles. A modern challenge for automotive engineers is to avoid these trust miscalibration issues through the development of techniques for measuring drivers' trust in the automated driving system during the execution of real-time applications. One possible approach for measuring trust is through modeling its dynamics and subsequently applying classical state estimation methods. This chapter proposes a framework for modeling the dynamics of drivers' trust in automated driving systems and also for estimating dynamic trust. The estimation method integrates sensed behaviors (from the driver) through a Kalman filter-based approach. The sensed behaviors include eye-tracking signals, the usage time of the system, and drivers' performance on a non-driving-related task (NDRT). A study ($n = 80$) with a simulated SAE level 3 automated driving system will be presented, and the factors that impact drivers' trust in the system will be analyzed. Data from the user study are used for the identification of the trust model parameters. Results will show that the proposed approach was successful in computing trust estimates over successive interactions between the driver and the automated driving system. These results encourage the use of strategies for modeling and estimating

trust in automated driving systems. This trust measurement technique paves a path for the design of trust-aware automated driving systems capable of changing their behaviors to control drivers' trust levels to mitigate both undertrust and overtrust.

This chapter is based on the work published in [6], and is organized in the following sections. Section 4.2 briefly describes the trust estimation problem and Section 4.3 focuses on the methodology that was applied for its solution. Section 4.4 details the user study conducted to validate the proposed method and the collected data from that study. Section 4.5 analyzes the collected data and presents the trust estimation results. Section 4.6 discusses the main contributions and implications of this new trust estimation framework and its limitations and, finally, Section 4.7 concludes the chapter.

4.2 Problem Statement

In this chapter, our main problem is to estimate drivers' trust in ADS from drivers' behaviors and actions in real-time while they operate a vehicle equipped with an SAE Level 3 ADS and concurrently perform a visually demanding NDRT. Our method must provide continuous trust estimates that can vary over time, capturing the dynamic nature of drivers' trust in the ADS. The estimation method must avoid the impractical process of repeatedly asking drivers their levels of trust in the ADS, and be as unobtrusive as possible for sensing drivers' behaviors and actions.

4.3 Method

4.3.1 Scope

To define the scope of our problem, we make the following assumptions about the ADS and the driving situation:

1. the ADS explicitly interacts with the driver in events that occur during vehi-

cle operation, and provides automated lane-keeping, cruise speed control, and collision avoidance capabilities to the vehicle;

2. the NDRT device is integrated with the ADS, allowing the ADS to monitor drivers' NDRT performance. The ADS can also track driver's head and eyes orientations;

3. drivers can alternate between using and not using the driving automation functions (i.e., the vehicle's self-driving capabilities) at any time during the operation;

4. when not using the driving automation functions, drivers have to perform the driving task, and therefore operate the vehicle in regular (non-automated) mode;

5. using the capabilities provided by the ADS, the vehicle autonomously drives itself when the road is free, but it is not able to maneuver around obstacles (i.e., abandoned vehicles) on the road. Instead, the ADS warns the driver whenever an obstacle is detected by the forward collision alarm system at a fair reaction distance. In these situations, drivers must take over driving control from the ADS and maneuver around the obstacle manually to avoid a collision; and

6. the forward collision alarm system is not perfectly reliable, meaning that both false alarms and misses can occur, and the ADS acknowledges when these errors occur. These false alarms and misses lead to interactions that are likely to decrease drivers' trust in the ADS. No other system malfunctions were implemented in the simulation.

4.3.2 Solution Approach

Assuming that the variations of trust caused by the interactions between the driver and the ADS can be quantified, we decide to apply a classical Kalman filter-based

continuous state estimation approach for trust. There are three reasons for applying a Kalman filter-based approach: (i) the fact that the continuous output measures of the estimator could be useful for the design of controllers and decision making algorithms in future applications; (ii) the aforementioned well-accepted practice of using continuous numerical estimates for trust in automated systems; and (iii) the difficulties related to the stochasticity of drivers' behaviors, which can be mitigated by the Kalman filter with recurring measurements. Therefore, to represent trust as a state variable, we need the mathematical derivation of a state-space model that represents the dynamics of trust. We assume that the dynamics of trust is influenced by the trustor agents' instantaneous level of trust and their experiences over time [14]. Those experiences are represented by interactions between the ADS and the driver associated with the reliability (or *internal risk*) of the ADS forward collision alarm. This assumption is an implication of Chapter III's conclusion that only internal risk affects ADS trust, while the external risk does not. Specifically, we consider that true alarms indicate high reliability and are positive experiences for the driver, while high internal risk manifestations given by false alarms and misses are negative driver experiences.

The implementation of a Kalman filter requires the definition of observation variables that can be measured and processed in real-time. These observation variables must be related to the variable to be estimated. Therefore, to satisfy the ease of implementation requirements stated in Section 4.2, we select a set of variables that were easy to sense and suitable for being used in a vehicular spatial configuration. The variables are: (i) the amount of time drivers spent using the autonomous capabilities provided by the ADS, i.e., *ADS usage time ratio*; (ii) the relative amount of time drivers spent focusing on a secondary task (the NDRT), measured with an eye-tracker device, i.e., *focus time ratio* [67]; and (iii) drivers' performance on that same NDRT, i.e., *NDRT performance*. The focus time ratio obtained with the eye

tracker is chosen because it is conveniently easy to be measured in a vehicle, and has been shown to be successfully representative of trust metrics [67]. The other variables are chosen because they are assumed to be proportional to trust: the more a driver trusts an ADS, the more s/he will use it; the more a driver trusts the ADS, the better s/he will perform on her/his NDRT.

Finally, to identify the parameters of a model for drivers' trust in ADS, we need to obtain a training dataset containing both inputs and their corresponding outputs. The outputs must be represented by drivers' true levels of trust in the ADS, which we can obtain by collecting their self-reports in a controlled user experiment. Therefore, only for the purpose of obtaining this training dataset, we establish a procedure for asking drivers their levels of trust in the ADS.

4.3.3 Definitions

To implement our solution methodology, we must first define the terms that will be used in our formulation.

Definition 1 (Trial)

A *trial* is concluded each time the driver operates the vehicle and reaches the end of a predefined route.

Trials are characterized by their time intervals, limited by the instants they start and end. Denoting these by t_0 and t_f, $t_0 < t_f$, the time interval of a trial is given by $[t_0, t_f] \in \mathbb{R}^+$.

Definition 2 (Event)

An *event*, indexed by $k \in \mathbb{N} \setminus \{0\}$, is characterized each time the ADS warns **or** fails to warn the driver about an obstacle on the road. Events occur at specific time instances t_k corresponding to k, $t_0 < \cdots < t_k < \cdots < t_f$, when the ADS:

1. correctly identifies an obstacle on the road and alerts the driver to take over control;

2. provides a false alarm to the driver; or

3. misses an existent obstacle and does not warn the driver about it.

Definition 3 (Event Signals)

The *event signals* are booleans $L(t_k)$, $F(t_k)$ and $M(t_k)$ corresponding to the event k that indicates whether the event was:

1. a true alarm, for which $L(t_k) = 1$ and $F(t_k) = M(t_k) = 0$;

2. a false alarm, for which $F(t_k) = 1$ and $L(t_k) = M(t_k) = 0$; or

3. a miss, for which $M(t_k) = 1$ and $L(t_k) = F(t_k) = 0$.

Definition 4 (Instantaneous Trust in ADS)

Drivers' *instantaneous trust in ADS* at the time instance t, $t_0 \leq t \leq t_f$ is a scalar quantity, denoted by $T(t)$.

$T(t)$ is computed from trust variation self-reports and from questionnaires answered by the driver, adapted from the work by Muir and Moray [83]. We re-scale the numerical range of the survey responses to constrain $T(t) \in [T_{min}, T_{max}]$, and arbitrarily choose $T_{min} = 0$ and $T_{max} = 100$. We also assume that $T(t)$ is immutable between two events, i.e., for $t_k \leq t < t_{k+1}$. We consider $T(t)$ to be our basis for the development of the proposed trust estimator.

Definition 5 (Instantaneous Estimate of Trust in ADS)

The *estimate of trust in ADS* at the time instance t, $t_0 \leq t \leq t_f$ is the output of the trust estimator to be proposed, and is represented by $\hat{T}(t)$. Its associated covariance is denoted by $\hat{\Sigma}_T(t)$.

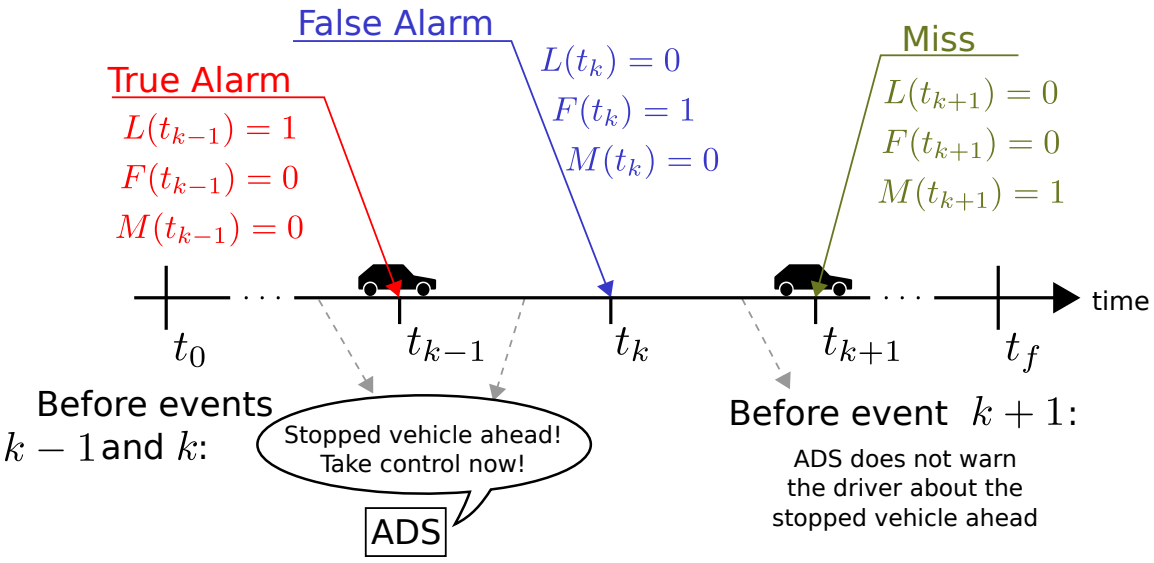

Figure 4.1: Timeline example for the stated problem. The event $k-1$ is a true alarm (there is an obstacle car and the ADS warns the driver about it); the event k is a false alarm (there is no car but the ADS also warns the driver); and the event $k+1$ is a miss (there is an obstacle car and the ADS does not warn the driver about it).

Definition 6 (Focus)

Drivers' *focus* on the NDRT, represented by $\varphi(t_k)$, is the percentage of time the driver spends looking at the NDRT screen during the interval $[t_k, t_{k+1})$.

Definition 7 (ADS Usage)

Drivers' *ADS usage*, represented by $\upsilon(t_k)$, is defined by the percentage of time the driver spends using the ADS self-driving capabilities during the interval $[t_k, t_{k+1})$.

Definition 8 (NDRT Performance)

Drivers' *NDRT performance*, represented by $\pi(t_k)$, is the total points obtained by the driver in the NDRT during the interval $[t_k, t_{k+1})$ divided by $\Delta t_k = t_{k+1} - t_k$.

We also call $\varphi(t_k)$, $\upsilon(t_k)$, and $\pi(t_k)$ our *observation variables*.

Figure 4.1 shows an example of a timeline scale that represents events within a trial. The NDRT and its score policies are explained in Section 4.4.

4.3.4 Trust Dynamics Model

To translate Castelfranchi's and Falcone's main aspects of trust dynamics [14] into mathematical terms, we must represent the experiences of the trustor agent, the subsequent change in trust, and relate those variables. Describing the user experiences with the passing time and the event signals, while also considering their discrete nature, we can expect a general relationship with the form represented by Equation (4.1),

$$T(t_{k+1}) = f(t_k, T(t_k), L(t_k), F(t_k), M(t_k)) \,, \tag{4.1}$$

where $f : [t_0, t_f] \times [T_{min}, T_{max}] \times \{0, 1\}^3 \to [T_{min}, T_{max}]$.

Additionally, we can expect the relationship between observations and trust to take the form represented by Equation (4.2),

$$\begin{bmatrix} \varphi(t_k) \\ \upsilon(t_k) \\ \pi(t_k) \end{bmatrix} = h(t_k, T(t_k), L(t_k), F(t_k), M(t_k)) \,, \tag{4.2}$$

where $h : [t_0, t_f] \times [T_{min}, T_{max}] \times \{0, 1\}^3 \to [0, 1]^2 \times \mathbb{R}$.

For simplicity, we assume the functions f and h to be linear, time-invariant, with additional random terms representing drivers' individual biases. Moreover, we model trust and the observation variables as Gaussian variables, and consider the observations to be independent of the event signals and within each other, representing the dynamics of trust in the ADS with the LTI system state-space model in Equations

(4.3),

$$\begin{cases} T(t_{k+1}) = \mathbf{A}T(t_k) + \mathbf{B}\begin{bmatrix} L(t_k) \\ F(t_k) \\ M(t_k) \end{bmatrix} + u(t_k)\,; \\ \begin{bmatrix} \varphi(t_k) \\ \upsilon(t_k) \\ \pi(t_k) \end{bmatrix} = \mathbf{C}T(t_k) + w(t_k)\,, \end{cases} \tag{4.3}$$

where $\mathbf{A} = \begin{bmatrix} a_{11} \end{bmatrix} \in \mathbb{R}^{1\times1}$, $\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \end{bmatrix} \in \mathbb{R}^{1\times3}$, $\mathbf{C} = \begin{bmatrix} c_{11} & c_{21} & c_{31} \end{bmatrix}^\top \in \mathbb{R}^{3\times1}$, $u(t_k) \sim \mathcal{N}(0, \sigma_u^2)$ and $w(t_k) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_w)$.

4.3.5 Trust Estimator Design

The state-space structure permits the application of Kalman filter-based techniques for the estimator design. We then propose the procedure presented in Algorithm 1. Figure 4.2 shows a block diagram representation of this framework, highlighting the trust estimator role in the interaction between the driver and the ADS.

4.4 User Study and Data Collection

We reproduced the situation characterized in Section 4.3 with the use of an ADS simulator. A total of 80 participants were recruited (aged 18-51, $M = 25.0$, $SD = 5.7$, 52 male, 26 female and 2 who preferred not to specify their genders). Participants were recruited via email and printed poster advertising. All regulatory ethical precautions were taken. The research was reviewed and approved by the University of Michigan's Institutional Review Board (IRB).

Algorithm 1 Trust Estimator

procedure TRUST_ESTIMATION($\hat{T}(t_k), \hat{\Sigma}_T(t_k)$, $L(t_k), F(t_k), M(t_k), \varphi(t_k), \upsilon(t_k), \pi(t_k)$)
  if $k = 0$ **then**
    $\hat{T}(t_0) \leftarrow (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top \begin{bmatrix} \varphi(t_0) \\ \upsilon(t_0) \\ \pi(t_0) \end{bmatrix}$
    $\hat{\Sigma}_T(t_0) \leftarrow 1$ ▷ Initializes trust estimate and co-variance
  else
    $K \leftarrow \hat{\Sigma}_T(t_k)\mathbf{C}^\top(\mathbf{C}\hat{\Sigma}_T(t_k)\mathbf{C}^\top + \Sigma_w)^{-1}$ ▷ Measurement update starting with Kalman gain computation
    $\begin{bmatrix} \hat{\varphi}(t_k) \\ \hat{\upsilon}(t_k) \\ \hat{\pi}(t_k) \end{bmatrix} \leftarrow \mathbf{C}\hat{T}(t_k)$
    $\mathbf{v} \leftarrow \begin{bmatrix} \varphi(t_k) \\ \upsilon(t_k) \\ \pi(t_k) \end{bmatrix} - \begin{bmatrix} \hat{\varphi}(t_k) \\ \hat{\upsilon}(t_k) \\ \hat{\pi}(t_k) \end{bmatrix}$ ▷ Innovation
    $T(t_k) \leftarrow \hat{T}(t_k) + K\mathbf{v}$
    $\Sigma_T(t_k) \leftarrow \hat{\Sigma}_T(t_k) - K\mathbf{C}\hat{\Sigma}_T(t_k)$
    $\hat{T}(t_{k+1}) \leftarrow \mathbf{A}T(t_k) + \mathbf{B} \begin{bmatrix} L(t_k) \\ F(t_k) \\ M(t_k) \end{bmatrix}$ ▷ Time Update
    $\hat{\Sigma}_T(t_{k+1}) \leftarrow \mathbf{A}\Sigma_T(t_k)\mathbf{A}^\top + \sigma_u$
  end if
  return $\hat{T}(t_{k+1}), \hat{\Sigma}_T(t_{k+1})$
end procedure

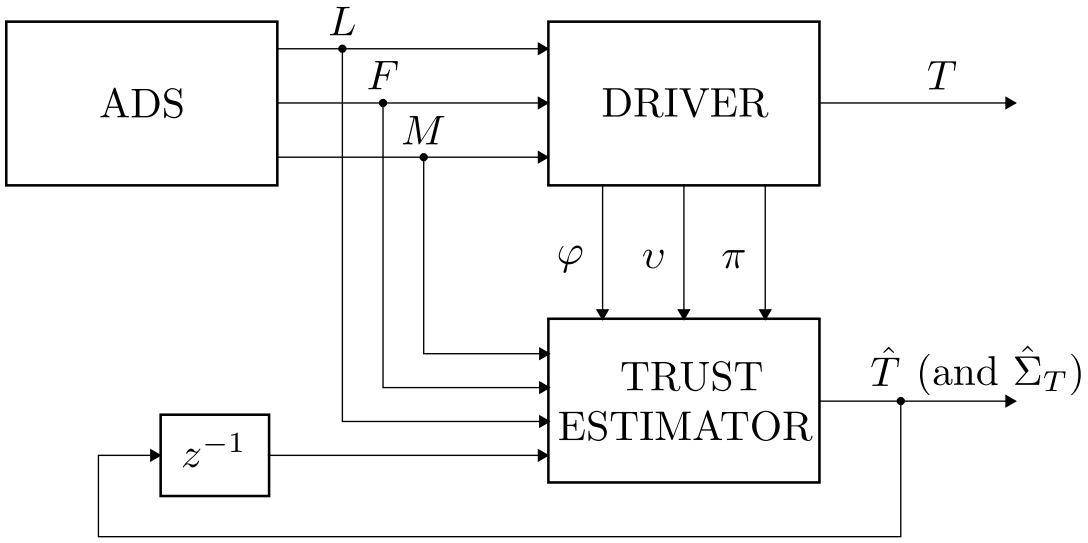

Figure 4.2: Block diagram representing the trust estimation framework. The event signals L, F, and M indicate the occurrence of a true alarm, a false alarm, or a miss. The observations φ, υ and π represent the drivers' behaviors. T is drivers' trust in ADS while $\hat{T}$ and $\hat{\Sigma}_T$ are the estimates of trust in ADS and the covariance of this estimate. A delay of one event is represented by the z^{-1} block.

4.4.1 Experiment and Data Collection

4.4.1.1 Study design

We employed a 4 (ADS error types) $\times$ 2 (road shapes) mixed user experimental design. Each participant experienced 2 trials, and each trial had 12 events. These 2 trials had the same ADS error type (between-subjects condition) and 2 different road shapes (within-subjects condition). The ADS error types that varied between subjects corresponded to 4 different conditions: control, for which all 12 events were true alarms; false alarms only, for which the 2^{nd}, 3^{rd}, 5^{th}, and 8^{th} events were false alarms; misses only, for which the 2^{nd}, 3^{rd}, 5^{th}, and 8^{th} events were misses; and false alarms and misses combined condition, for which the 2^{nd} and 5^{th} events were false alarms, while the 3^{rd} and 8^{th} events were misses. The ADS error type was assigned according to the participants' sequential identification number. The road shapes were represented by straight and curvy roads, and were assigned in alternating order to

minimize learning and ordering effects.

4.4.1.2 Tasks

The experimental setup was very similar to the one described in Chapter III, the main difference was the layouts of the roads. We used the ANVEL simulator [30] and the NDRT was the previously used adapted version of the Surrogate Reference Task [47], implemented with PEBL [80]. Figure 4.3 (a) shows the experimental setup with the tasks performed by the driver.

In the driving task, participants operated a simulated vehicle equipped with an ADS that provided it automatic lane keeping, cruise control, and collision avoidance features. Participants were able to activate the ADS (starting autonomous driving mode) by pressing a button on the steering wheel, and to take back control by braking or by steering. Figure 4.3(b) shows the driving task interface with the driver.

With the ADS activated (i.e., with the vehicle in self-driving mode), participants were expected to execute the visual search NDRT. They were not allowed to engage in both driving and executing the NDRT simultaneously, and the experimenters would stop the test if they did so. Participants were informed that the vehicle could request their intervention if they identified obstacles on the road, as it is expected for Level 3 ADSs [101]. Figure 4.3(c) shows the NDRT interface with the driver.

Participants could not focus only on the NDRT, because the ADS demanded them to occasionally take control of the driving task. They were asked to be ready to take control upon intervention requests from the ADS, as some obstacles occasionally appeared on the road. At that point, the ADS identified the obstacles and asked the driver to take control, as the vehicle was not able to autonomously change lanes and maneuver around them. If drivers did not take control, the emergency brake was triggered when the vehicle got too close to an obstacle, and then drivers lost points on their ongoing NDRT score. In that situation, they still needed to take control of the

driving task, maneuver around the obstacle and re-engage the autonomous driving mode. They obtained 1 point for each correctly chosen "Q" and lost 5 points each time the emergency brake got triggered.

With the events characterized by true alarms or misses, drivers had to take control and pass the obstacle. Subsequently, they were asked about their "trust change". When asked, they had to stop the vehicle to answer the question on a separate touchscreen. They reported their trust change in the events characterized by true alarms, false alarms, and misses. They had 5 choices, varying from "Decreased Significantly" to "Increased Significantly", as shown in Figure 4.3(d). These choices were then used as indicators of the differences $\Delta T_k^Q \in \{-2, -1, 0, 1, 2\}$ (we use the superscript Q to indicate that the differences were quantized).

4.4.1.3 Procedure

Upon arrival, participants were asked to complete a consent form as well as a pre-experiment survey related to their personal information, experience with ADS, mood and propensity to trust the ADS. After the survey, the tasks were explained and the experimenter gave details about the experiment and the simulated vehicle control. Participants then completed a training session before the actual experiment began and, in sequence, completed their two trials. After each trial, participants were asked to complete post-trial surveys related to their trust in the ADS. These surveys were administered electronically. Each trial took approximately 10 to 15 minutes, and the whole experiment lasted approximately 60 minutes.

A basic fixed level of cash compensation of $15.00 was granted for the participants. However, they also had the possibility of receiving a performance bonus. The bonus was calculated according to their best final NDRT score, considering both trials experienced by the participant. Those who made up to 199 points in the NDRT did not receive a bonus. However, bonuses of $5.00 were granted for those who made

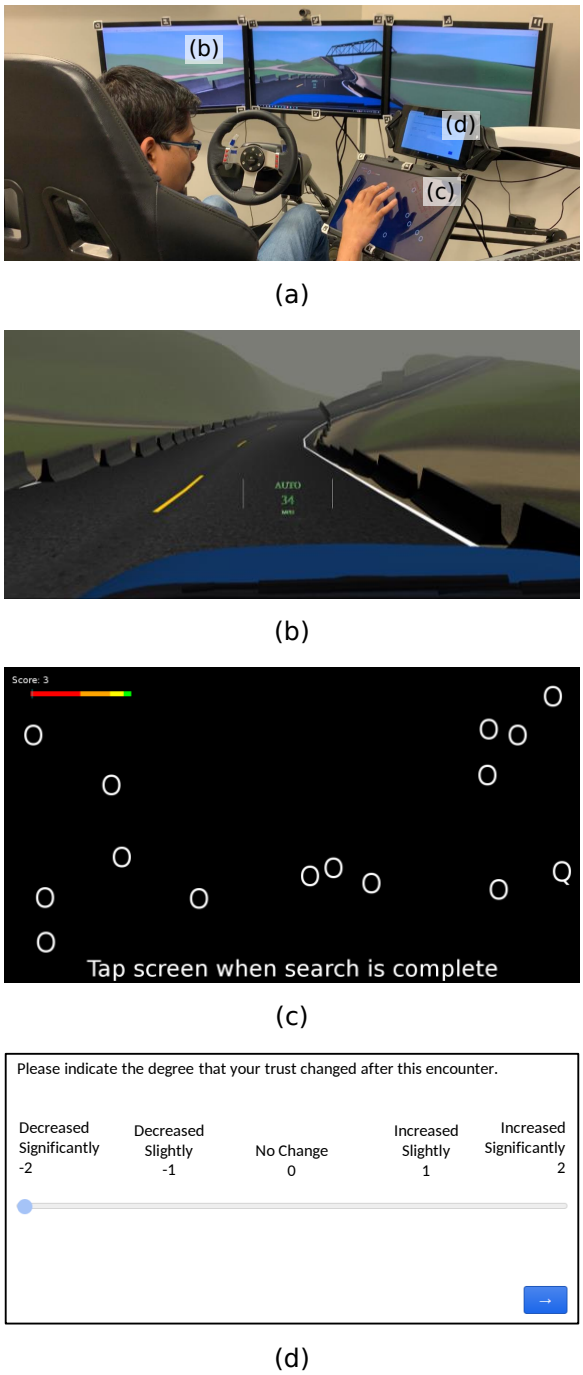

Figure 4.3: Experimental design (a), composed of the driving task (b), the NDRT (c) and the trust change self-report question (d). The trust change self-report question popped up after every event within the trials (there were 12 events per trial), including true alarms, false alarms, and misses.

between 200 and 229 points; $15.00 for those who made between 230 and 249 points; and $35.00 for those who made 250 points or more. From the total of 80 participants,

28 got $5.00 bonuses, 6 participants got $15.00 bonuses, and no participant got the $35.00 bonus.

4.4.1.4 Apparatus

As illustrated in Figure 4.3(a), the simulator setup was composed of three LCD monitors integrated with a *Logitech* G-27 driving kit. Two other smaller touchscreen monitors positioned to the right hand of the participants were used for the NDRT and for the trust change self-report questions. The console was placed to face the central monitoring screen so as to create a driving experience as close as possible to that of a real car. In addition, we used *Pupil Lab's Pupil Core* eye tracker mobile headset, equipped with a fixed "world camera" to measure participants' gaze positional data.

4.4.1.5 Measured Variables

Measured variables included participants' subjective responses, behavioral responses and performance. Observation variables $\varphi(t_k)$, $\upsilon(t_k)$ and $\pi(t_k)$ were also measured and averaged for the intervals $[t_k, t_{k+1}]$. Subjective data was gathered through surveys before and after each trial, including trust perception, risk perception, and workload perception. We used questionnaires adapted from [83] and [100] to measure post-trial trust and risk perception, respectively. Eye-tracking data included eyes' positions and orientations, as well as videos of the participants' fields of view.

$T(t_k)$ was computed from the post-trial trust perception self-reports $T(t_f)$ and the within trial trust change self-reports ΔT_k^Q, as in Equation (4.4),

$$\begin{cases} T(t_{12}) = T(t_f)\,; \\ T(t_k) = T(t_f) - \alpha \sum_{i=k+1}^{12} \Delta T_i^Q\,, \end{cases} \tag{4.4}$$

where $k \in \{0, 1, 2, \ldots, 11\}$, and $\alpha = 3$. Therefore, the trust measures $T(t_k)$ were

back-computed for the events within a trial. The α value was chosen to characterize noticeable variations in $T(t_k)$, but also avoiding $T(t_k)$ values falling outside the interval $[T_{min}, T_{max}]$. Positive values for α between 1 and 3 were tested and provided results similar to those reported in Section 4.5.

4.4.2 Model Parameters

Considering the formulation presented in Section 4.3 and the data obtained in the user study, we turn to the identification of parameters for the trust model and the design of the trust estimator. We found the best fit parameters for the short-term (i.e., with respect to events) trust dynamics represented by the state-space model in Equation (4.3). From the 80 participants, we selected 4 from the dataset—each one chosen randomly within each of the 4 possible ADS error type conditions—and used the data from the remaining 76 to compute the parameters, which are presented in Table 4.1. We used the data from the 4 selected participants for validation. The parameters of the state-space model from Equation (4.3) were identified with maximum likelihood estimation through linear mixed-effects models. Our models included a random offset per participant to capture their individual biases and mitigate the effects of these biases in the results, and to represent normally distributed random noises.

4.5 Results

4.5.1 Participants' Data Analysis

For each of the observation variables, we obtained 1920 measurements (80 participants $\times$ 2 trials per participant $\times$ 12 events per trial). The parameters describing these distributions are presented in Table 4.2. The histograms for these distributions are shown in Figure 4.4; the probability density functions corresponding to normal

Table 4.1: Trust in ADS state-space model parameters

Parameter	Value Estimate	S.E.M†
a_{11}	0.9809	4.0×10^{-3}
b_{11}	3.36	0.29
b_{12}	-0.61	0.32
b_{13}	-1.30	0.31
c_{11}	6.87×10^{-3}	3.3×10^{-4}
c_{21}	9.10×10^{-3}	1.0×10^{-4}
c_{31}	4.38×10^{-3}	1.0×10^{-4}
σ_u^2	1.24	–
$\mathbf{\Sigma}_w$	$\text{diag}(1.0, 1.6, 1.8) \times 10^{-3}$	–

†S.E.M = Standard error of the mean.

distributions $\mathcal{N}(\mu_\varphi, \sigma_\varphi^2)$, $\mathcal{N}(\mu_\upsilon, \sigma_\upsilon^2)$ and $\mathcal{N}(\mu_\pi, \sigma_\pi^2)$ are also shown.

Table 4.2: Parameters for the Focus φ, ADS usage υ and NDRT performance π measurements distributions

Parameter	Distributions		
	φ	υ	π
Minimum	0.02	0.17	0.00
25^{th} percentile	0.32	0.69	0.28
50^{th} percentile	0.47	0.74	0.33
75^{th} percentile	0.65	0.79	0.38
Maximum	0.97	0.92	0.56
Mean μ	0.49	0.73	0.32
Standard Deviation σ	0.20	0.08	0.08

The plots in Figure 4.5 present the average trust over interactions for all participants in each ADS error type conditions, indicating the occurrence of true alarms, false alarms and misses (represented by 'T', 'F' and 'M', respectively). The curves are consistent with the expected behavior for the state-space model (4.3) and the model parameters given in Table 4.1. These plots are similar to those presented in Figure

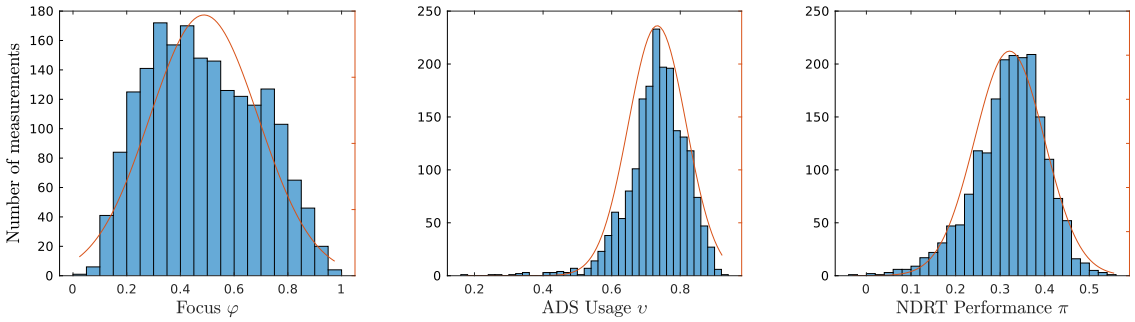

Figure 4.4: Histograms for the Focus φ, ADS usage υ and NDRT performance π measurements distributions and overlapping probability density functions with corresponding means and standard deviations. Each distribution had 1920 measurements (= 80 participants $\times$ 2 trials per participant $\times$ 12 measurements per trial).

3.7, with the difference that the reliability of the ADS was manipulated not only with false alarms, but also with misses.

4.5.2 Trust Estimation Results

After obtaining the model parameters, we applied Algorithm 1 to estimate the trust levels of the participants that were excluded from the dataset. Figure 4.6(a1:a4) and Figure 4.7(a1:a4) present the trust estimation results for these participants (identified as A, B, C and D). Participant A experienced the combined ADS error type condition; participant B experienced the false alarms only condition; participant C experienced the control condition; and participant D experienced the misses only condition. The plots bring together their two trials and the different estimate results for each trial. For participants A and B, trial 1 was conducted on a curvy road and trial 2 on a straight road. For participants C and D, trial 1 was conducted on a straight road and trial 2 on a curvy road.

The accuracy of our estimates improved over time as the participants interacted with the ADS. Figure 4.6(a1) shows that, for participant A, trial 1, the initial trust estimate $\hat{T}(t_0)$ and the initial observed trust $T(t_0)$ were close to each other (in comparison to Figure 4.6(a2)). This means that the estimate computed from the observations

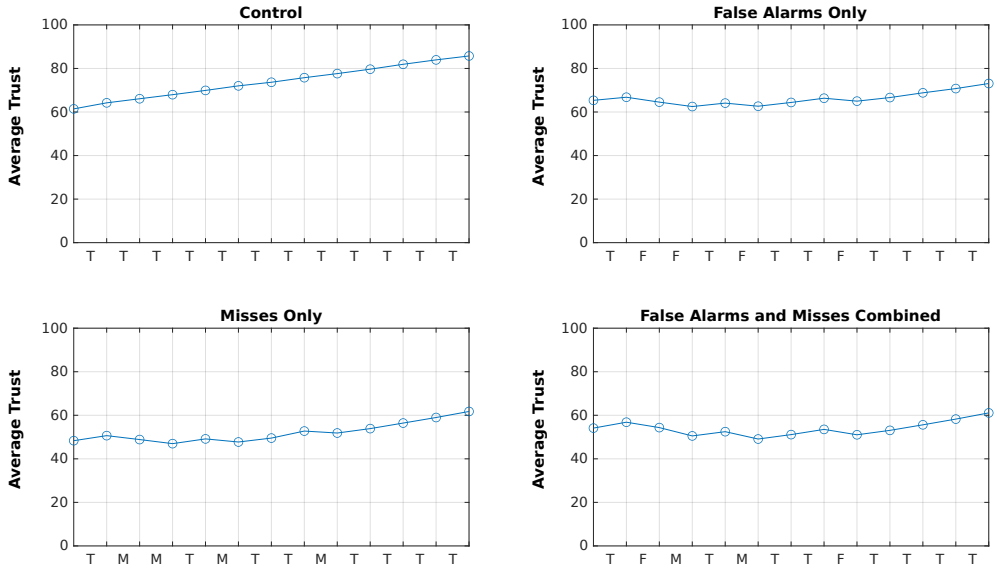

Figure 4.5: Plots of the average trust for all participants in each ADS error type condition. When participants were using a reliable ADS, i.e., in the *Control* condition, trust increased steadily after the true alarms indicated by 'T' in the horizontal axes. After false alarms or misses (indicated respectively by 'F' and 'M') occurred, trust decreased accordingly.

taken at the beginning of the trial, i.e., $\varphi(t_0)$, $\upsilon(t_0)$, and $\pi(t_0)$, approximately matched the participants' self-reported trust level. Considering the Kalman filter's behavior, the curves remained relatively close together over the events, as expected. Therefore the estimate followed the participants' trust over the trial events. This accuracy, however, was not achieved at the beginning of the second trial, as can be observed in Figure 4.6(a2). This figure shows that, in trial 2, $\hat{T}(t_0)$ and $T(t_0)$ had a greater difference, but this difference decreased over the events as the curves converged. A similar effect can be observed for participants B, trial 2 as in Figure 4.6(a3:a4) and for participant C, as in Figure 4.7(a1:a2).

Participants' responses to similar inputs were not always coherent, and varied over time or under certain conditions. Predominantly, participants' self-reported trust increased after true alarms (indicated by the prevailing positive steps at the events that are characterized by orange circles). In addition, after false alarms and misses,

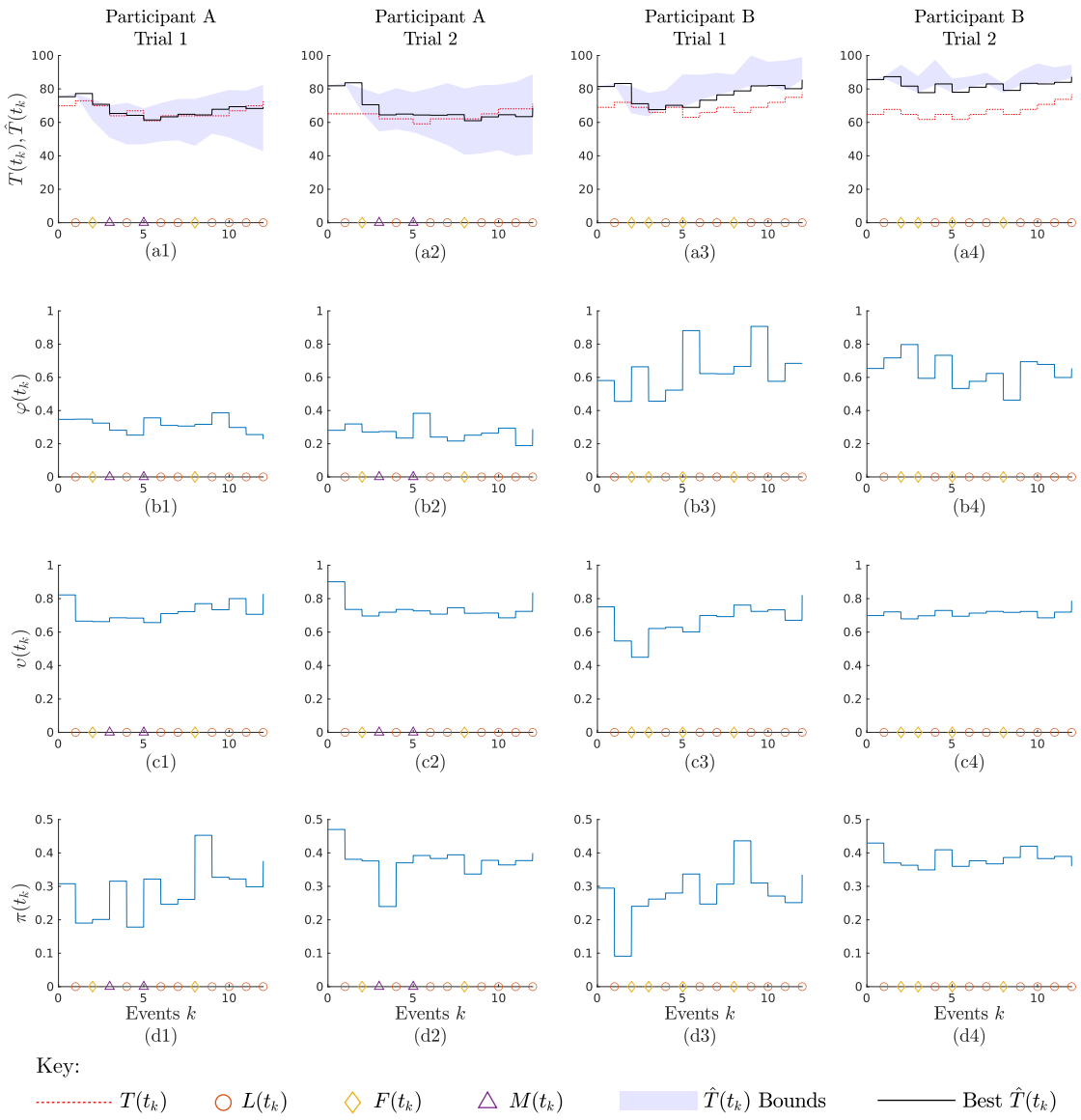

Figure 4.6: Trust estimation results for participants A and B. Participant A experienced both false alarms and misses (combined ADS error type condition) while participant B experienced false alarms only (false alarms only condition). For both participants, the first trial was conducted on a curvy road, while the second trial was conducted on a straight road. Curves in (a1:a4) show the estimation results, indicating that the estimator can track the trust self-reports, i.e., $\hat{T}(t_k)$ approaches $T(t_k)$ over the events. This is made possible with the processing of the observations variables focus time ratio (φ), ADS usage time ratio (υ), and NDRT performance (π) presented in (b1:d4).

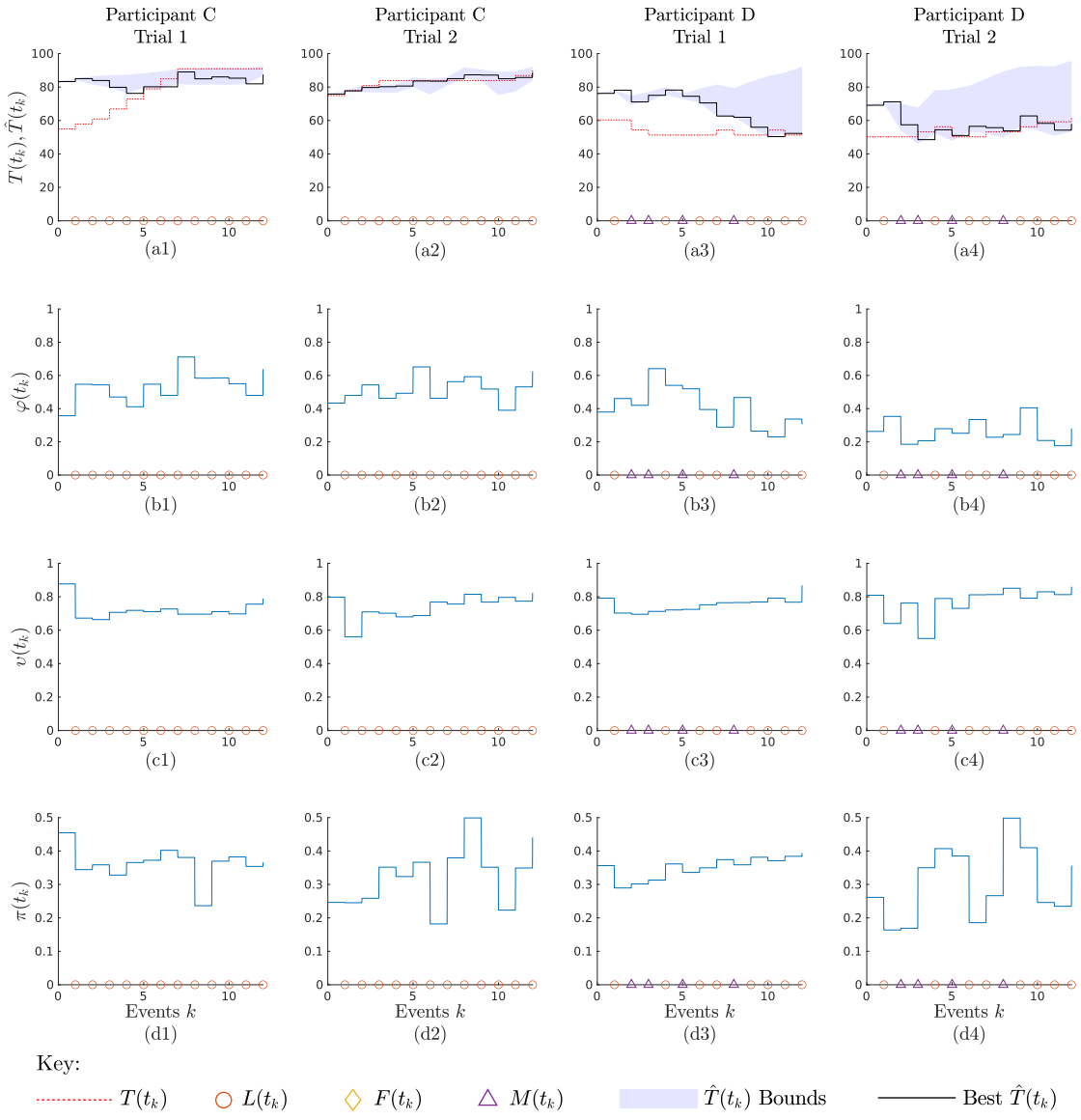

Figure 4.7: Trust estimation results for participants C and D. Participant C experienced only true alarms (control ADS error type condition) while participant D experienced misses only (misses only condition). For both participants, the first trial was conducted on a straight road, while the second trial was conducted on a curvy road.

they usually reported trust decreases (indicated by the prevailing negative steps at the events characterized by yellow diamonds and purple triangles). However, it is noticeable that, for participant A, trial 2, the self-reported trust was more "stable", as indicated by fewer steps on the red dashed curve. Two different factors could have contributed to the less frequent variations on $T(t_k)$: as the participant was on a straight road, the perceived risk might not have been high enough to induce drops after false alarms; or, as it was the participant's second trial, the learning effects might have softened the self-reported trust changes (especially after false alarms). In any case, the difference between the curve patterns in Figure 4.6(a1) and Figure 4.6(a2) suggests a non-constancy on participant A's characteristic behaviors. A similar behavior was observed for participant C, trial 1 after the 8^{th} alarm and for trial 2.

The observation variables we selected were effective in representing drivers' trusting behaviors. Figure 4.6(b1:d4) show the observation variables corresponding to the trust curves in Figure 4.6(a1:a4), while Figure 4.7(b1:d4) correspond to 4.7(a1:a4). All observation variables have a positive correlation with trust, and therefore it can be observed that some noticeable peaks and drops in the observation variables correspond to positive and negative variations in the estimate of trust in ADS. This is especially true for the counter-intuitive behaviors of the participants. For instance, as it can be seen in Figure 4.6(a3:d3), after the 8^{th} event—which was a false alarm—participant B reported a drop in his/her trust level, indicating that $T(t_8) < T(t_7)$. However, his/her behaviors did not reflect that drop: we can notice that $\varphi(t_8) > \varphi(t_7)$, $\upsilon(t_8) > \upsilon(t_7)$ and $\pi(t_8) > \pi(t_7)$. As a result, the trust estimate had an increase, and eventually we had $\hat{T}(t_8) > \hat{T}(t_7)$. Similar counter-intuitive situations can be identified for participants A, C and D.

The accuracy of the estimates depends on the covariance parameters, which can be tailored for the driver. The trust estimate bounds represented by blue bands in Figure 4.6(a1:a4) and Figure 4.7(a1:a4) are approximations obtained with the over-

lay of several simulations (100 in total). This variability is due to the uncertainty represented by the random noise parameters $u(t_k)$ and $w(t_k)$, and the width of the bound bands is related to the computed covariances σ_u^2 and $\mathbf{\Sigma}_w$. Both lower values for σ_u^2 and higher values for $\mathbf{\Sigma}_w$ entries would imply a narrower band, meaning that the estimator would have less variability (and therefore could be slower on tracking trust self-reports). Meanwhile, higher σ_u^2 and lower values of $\mathbf{\Sigma}_w$ entries would imply, respectively, a less accurate process model and on observations considered more reliable. This would characterize wider bands, and thus the variations on the estimate curves would be more pronounced.

Trust estimates may be more accurate with the individualization of the model parameters. Although we used the average parameters presented in Table 4.1 for the results, a comparison of Figure 4.6(a2), Figure 4.7(a1) and Figure 4.7(a3:a4) with Figure 4.6(a4), suggests that the balance between σ_u^2 and $\mathbf{\Sigma}_w$ should be adapted to each individual driver. It can be seen that these parameters permitted a quick convergence of $T(t_k)$ and $\hat{T}(t_k)$ for participants A, C and D, but that 12 events were not enough for the estimator to track the trust self-reports from participant B. We also computed the root-mean-square (RMS) error of the estimate curves resulting from the 100 simulations for participants A, B, C and D. The RMS error distributions had the characteristics presented in Table 4.3.

Considering the 100-points trust range, for participant A the error stands below 10%, while for participants B, C and D it stands below 20%. This difference suggests that the parameters of the model are more suitable for participant A than for participant B, C and D.

Table 4.3: RMS error of the estimate curves from Figure 4.6 and Figure 4.7

Participant	Trial	Mean	Standard Deviation
A	1	4.9	2.4
A	2	10.0	2.1
B	1	14.5	2.8
B	2	19.1	1.2
C	1	14.2	0.4
C	2	2.7	0.6
D	1	20.7	2.2
D	2	13.8	3.4

4.6 Discussion

4.6.1 Implications

The goal of this chapter was to propose a framework for real-time estimation of drivers' trust in ADS based on drivers' behaviors and dynamic trust models. As shown by the results, our framework successfully provides estimates of drivers' trust in ADS that increase in accuracy over time. This framework is based on a novel methodology that has considerable advantages over previously reported approaches, mainly related to our trust dynamics model and the simpler methods needed for its implementation.

First, the sensing machinery required for implementing our methodology is as simple and as unobtrusive as possible. Considering practical aspects related to the framework implementation, we have chosen observation variables that are suitable for the estimation of drivers' trust in ADS. An eventual implementation of the proposed estimator on an actual self-driving vehicle would depend only on the utilization of an eye-tracking system and on the integration between the ADS and the tasks performed by the driver. Our unique observation variable that comes from a direct instrumentation of drivers' behavioral patterns is the eye-tracking-based focus on the NDRT.

The other observation variables (NDRT performance and ADS usage) are indirectly measured by the ADS. Eye-tracking-based metrics are appropriate for trust measuring as they do not require sensory devices that would be impractical and/or intrusive for drivers. Although we have used an eye tracker device that has to be directly worn by the participant, there exist different eye-tracking systems that do not need to get in direct contact with the driver to sense their gaze orientations, and could be used in a real world implementation of this framework.

Second, the results of our framework show that it can successfully estimate drivers' trust in ADS levels, but the accuracy of the estimates were different depending on the driver. The application of the model represented by Equation (4.3) in the trust estimator algorithm required average (population-wise) state-space model parameters. These parameters were computed with a minimization approach, and they are indications of reasonable statistics for average values conditioned to our pool of participants. However, these parameters could vary drastically from driver to driver. In a more sophisticated implementation of our modeling and estimation methodology, the values from Table 4.1 should serve as preliminary parameters only. A possible way to improve our proposed methodology would be to integrate it with learning algorithms to adapt the model parameters to individual drivers. Moreover, as drivers become accustomed to the ADS's operation, these parameters might also vary over time (making the time-invariant description from Equation (4.3) not useful). Therefore, an eventual ADS featuring our framework should also be sufficiently flexible to track the changes in individual drivers' model parameters over time, as proposed in [130].

Third, the framework opens paths for more research on the development of more complex models and estimation techniques for trust. These techniques may encompass both the driver-ADS context and other contexts characterized by the interaction between humans and robots. In the case of driver-ADS contexts, the events that

trigger the propagation of the trust state do not need to be restricted to the forward collision alarm interactions characterized by true alarms, false alarms and misses. A wider range of experiences could be considered in the process model represented by Equation (4.3), such as events related to the ADS driving performance or to external risk perceived by the ADS. Drivers could be engaged in alternative NDRTs, as long as they are integrated with the ADS and a continuous performance metric is defined as observation variable. In the case of interactions between humans and robots in different scenarios, the concepts that were defined in Section 4.3 are easily expandable to other contexts. The main requirement would be the characterization of what are the events that represent important (positive and negative) experiences within interactions between the human and robot. These positive and negative experiences would generally characterize the robot's performance, which is an essential factor describing the basis of trust, as identified by Lee and See [60]. Robots that execute specific tasks in goal-oriented contexts could have their performances measured in sequential time instances that would trigger the transition of the trust state. For instance, these performance measures could be a success/failure classification, such as pick and place task with a robotic arm [114, 122, 133]; or a continuous performance evaluation, such as when a follower robot loses track of its leader due to the accumulation of sensor error [102, 103].

Finally, the framework provides trust estimates that are useful for the design of trust controllers to be embedded in new ADSs. In this framework, trust is modeled as a continuous state variable, which is consistent with widely used trust scales and facilitates the processing and analysis of trust variations over time. This trust representation permits considering the incremental characteristics of the trust development phenomena, which is consistent with the literature on trust in automation and opens a path for the development of future trust control frameworks in ADSs. Since it is developed in the state-space form, our method for modeling drivers' trust

in ADS enables the use of classical application-proven techniques such as the Kalman filter-based method we have used in Algorithm 1.

In addition, a practical implication of the proposed estimation framework is that it could be used in innovative adaptive systems capable of estimating drivers' trust levels and reacting in accordance with the estimates, in order to control drivers' trust in ADS. These functionalities would need to involve strategies to monitor not only drivers' behaviors but also the reliability of the system (for example, the acknowledgment of false alarms and misses mentioned in Section 4.3.1, assumption 6.). These errors could be identified after a sequence of confirmations or contradictions of the sensors' states, while the vehicle gets closer to the event position, entering the ranges of higher accuracy of those sensors. Moreover, the system could request the driver to provide it feedback about issued alarms to identify its own errors, asking confirmation about identified obstacles or enabling quick report of missed obstacles, a functionality that is currently present in GPS navigation mobile applications [123]. Although these questions could represent an inconvenient distraction, this strategy is not as disruptive as demanding drivers to provide trust self-reports, especially during autonomous operation. The integration between the ADS and the NDRTs would also be needed for the assessment of observation variables and, eventually, actions to increase or decrease trust in ADS could be taken to avoid trust-related issues (such as under- and over-trust). These trust control schemes would be useful for improving driver-ADS interactions, having the goal of optimizing the safety and the performance of the team formed by the driver and the vehicle.

4.6.2 Limitations

4.6.2.1 Trust Modeling and Estimation Methodology

A limitation of the study presented in this chapter relates to the assumptions associated with how we derive the state-space model for trust in the ADS. The re-

lationships represented by Equations (4.1) and (4.2) restrict the experiences of the trustor agent (the driver) to the events represented by true alarms, false alarms and misses of the forward collision alarm. In fact, other experiences such as the ADS's continuous driving performances can characterize events that could be represented by signals of different types other than booleans. The simplification of the relationships represented by (4.1) and (4.2) to the LTI system represented by (4.3) is useful and convenient for the system identification process and for the trust estimator design. However, the resulting model fails to capture some phenomena that are likely to occur during the interactions between drivers and ADSs. These phenomena might include the variation of model parameters over time (i.e., after a reasonable period of drivers' interaction with the ADS) or the possibly nonlinear relationship between trust and the observation variables. An example is the relationship between trust and NDRT performance: it is unlikely that in a more rigorous modeling approach we could consider these variables to be directly proportional. Usually, an excess of trust (overtrust) in a system can lead to human errors, which might eventually result in performance drops.

4.6.2.2 User Study

There are several other limitations that relate to the experimental study presented in this chapter.

First, most participants were young students, very experienced with video games and other similar technologies. Our results could have been biased by these demographic characteristics.

Second, we employed a simulator in our experimental study. The use of a simulated driving environment is a means of testing potentially dangerous technologies. In general, people tend to act similarly in real and simulated environments [42]. However, due to the risks involved in driving, we acknowledge that participants might not have

felt as vulnerable as they would if this study had been conducted in a real car.

Finally, we employed a specific NDRT to increase the participants' cognitive load. The recursive visual search task gives drivers the opportunity to switch their attention between the driving and the NDRT very frequently. Other types of NDRTs could demand drivers' attention for longer periods of time, and this could induce a different effect on trust, risk perception or performance. The NDRT performance metric in this study is very specific and may or may not be generalizable to other task types.

4.6.3 Improvements and Usability

Additional improvements to our framework may be achieved by addressing the limitations of the reported user study. A vehicle with autonomous capabilities can be utilized to make the participants' experience as similar as possible to a realistic situation. Additionally, our methodology could be tested in other different scenarios where the complexity of the NDRT and of the environment are increased.

4.7 Conclusion and Contribution

The main contribution of this chapter is the proposed framework for the estimation of drivers' trust in ADSs. This framework is applicable for SAE level 3 ADSs, where drivers conditionally share driving control with the system, and that system is integrated with a visually demanding NDRT. In comparison to previous trust estimation approaches, it has practical advantages in terms of implementation ease and of the format of its trust estimates outputs.

We investigated the effectiveness of the proposed framework with a user study that is described in Section 4.4. In this user study, participants operated a simulated vehicle featuring an ADS that provided self-driving capabilities for the vehicle. Participants conducted two concurrent (driving and non-driving) tasks, while reporting their levels of trust in the ADS. Our goal was to establish a computational model for

drivers' trust in ADS that permitted trust prediction during the interactions between drivers and ADSs, considering the behaviors of both the system and the driver. We found the parameters of a discrete-time, LTI state-space model for trust in ADS. These parameters represented the average characteristics of our drivers, considering the resultant experiment dataset. With the calculation of the parameters, it was possible to establish a real-time trust estimator, which could track the trust levels over the interactions between the drivers and the ADS.

In summary, our results reveal that our framework was effective for estimating drivers' trust in ADS through the integration of the NDRT and behavioral sensors to ADSs. We also show, however, that a more advanced strategy for trust estimation must take into consideration the individual characteristics of the drivers, making systems flexible enough to adjust their model parameters during continuous use. Our technique opens ways for the design of smart ADSs able to monitor and dynamically adapt their behaviors to the driver, in order control drivers' trust levels and improve driver-ADS teaming. More accurate trust models can improve the performance of the proposed trust estimation framework and, therefore, are still required. However, the utilization of this trust estimation framework can be a first step to designing systems that can, eventually, increase safety and optimize joint performances during the interactions between drivers and ADSs embedded in self-driving vehicles.

The modeling technique and the trust estimator presented in this chapter could be used in the design of a trust management system. This trust management system could be based on the comparison of trust level estimates with the assessed capability and reliability of the vehicle in different situations, which depends on the risk involved in the operation. From the comparison, the trust calibration status could be evaluated, and a possible mismatch between trust and capability (or reliability) levels would indicate the need for system reaction. This reaction would consist of actions to manipulate trust levels, seeking to increase trust in case of distrust (or undertrust)

and to decrease it in case of overtrust. An example of trust management system with these characteristics is presented in Chapter V.

CHAPTER V

Calibration of Drivers' Trust in ADSs

5.1 Introduction

Automated vehicles (AVs) that intelligently interact with drivers must build a trustworthy relationship with them. A calibrated level of trust is fundamental for the AV and the driver to collaborate as a team. Techniques that allow AVs to perceive drivers' trust from drivers' behaviors and react accordingly are, therefore, needed for context-aware systems designed to avoid trust miscalibrations. This chapter proposes a framework for the management of drivers' trust in AVs. The framework is based on the identification of trust miscalibrations and on the activation of different communication styles to encourage or warn the driver when necessary. Our results show that the management framework is effective, increasing (decreasing) trust of undertrusting (overtrusting) drivers, and reducing the average trust miscalibration time periods by approximately 40%. Similar to the trust estimator proposed in Chapter IV, the trust management framework is applicable for the design of SAE Level 3 automated driving systems and has the potential to improve performance and safety of driver–AV teams.

This chapter is based on the work published in [7]. The remainder of the chapter is organized as follows. Section 5.2 presents the problem of identifying trust miscalibrations and manipulating drivers' trust in the AV (or, interchangeably, the ADS) to eventually achieve trust calibration. Section 5.2 also proposes a solution for that

problem, managing trust by combining the trust estimation method presented in Chapter IV with a rule-based controller to calibrate trust. Section 5.3 focuses on the implementation of the methods and the user study conducted to validate the trust management solution. Section 5.4 details the results obtained with the utilization of the trust management framework, and compares metrics of overall trust calibration between groups of participants that used and that did not use the proposed trust management framework. Section 5.6 concludes and presents a brief discussion on future directions for the research presented in this chapter.

5.2 Problem Statement

Considering the context of a driver interacting with an AV featuring an SAE Level 3 automated driving system (ADS), we addressed two main problems. First, we aimed to identify instances for which drivers' trust in the AV is miscalibrated, i.e., when the driver is undertrusting or overtrusting the AV. Second, we focused on manipulating drivers' trust in the AV to achieve calibrated levels, i.e., trust levels that match the AV's capabilities [60]. In other words, our goal was to increase or decrease drivers' trust in the AV whenever drivers were undertrusting or overtrusting the AV, respectively.

In SAE Level 3 ADSs, drivers are required to take back control when the system requests intervention or when it fails [101]. We assume that the AV has automated lane-keeping, cruise control and forward collision alarm functions that can be activated (all at once) and deactivated at any time by the driver. The AV can also identify different road difficulty levels and process drivers' behavioral signals to estimate their trust in the AV.

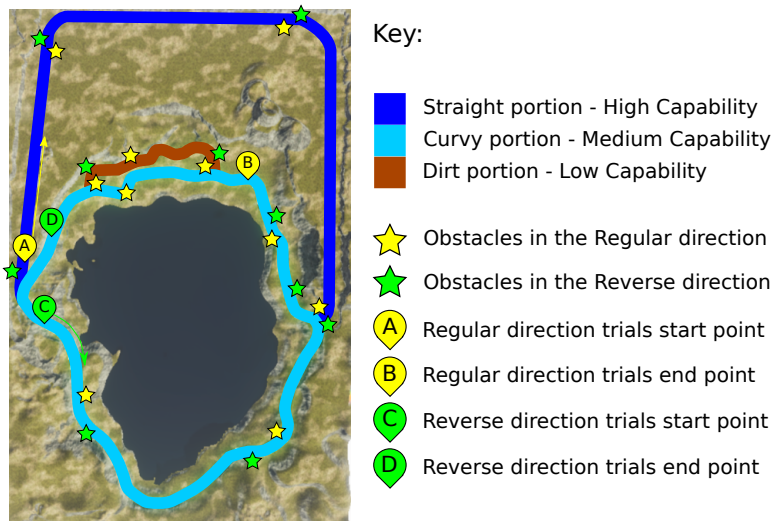

Figure 5.1: Circuit track used in this study. The portions of the road correspond to the capability of the AV. In the regular direction, drivers start at point A, follow the "straight" path in the clockwise direction, cover the curvy path and finish the trial at point B, right after passing through the dirt road portion. In the reverse direction, drivers start at point C, follow the curvy path in the counterclockwise direction, cover the straight path, continue to the curvy path (until the dirt portion), pass through the dirt portion, and finish the trial at point D. Both directions have 12 events (encounters with obstacles), and it took drivers approximately 10 to 12 minutes to complete a trial.

5.2.1 Solution Approach

We implemented a scenario to represent the described problem context with an AV simulator. We established simulations where drivers took trials in a predefined circuit track. The circuit track was divided into distinct parts, having three predefined risk levels, corresponding to the difficulty associated with each part of the circuit track. The easy parts of the circuit track consisted of predominantly straight roads; the intermediate difficulty parts were curvy paved roads; and the difficult parts were curvy dirt roads. Within these trials, drivers encountered abandoned vehicles on the road, which represented obstacles that the AV was not able to maneuver around by itself (using its automated driving functions). At that point, drivers had to take over control, pass the obstacle and then engage the autonomous driving mode again. Figure 5.1 shows the circuit implemented in the simulation environment.

We needed to compare drivers' trust levels and the AV's capability levels to identify trust miscalibrations. Therefore, we defined three capability levels for the AV, corresponding to the difficulty of the circuit track parts. The AV's forward collision alarm was able to identify the obstacles and also to trigger an emergency brake if the driver did not take control in time to maneuver around the obstacles. These two actions were activated at different distances to the obstacles, represented by the two circular regions represented in Figure 5.2. On straight paved roads these distances were larger, representing the longer perception ranges of the AV sensors. On more difficult parts of the circuit (i.e., curvy or dirt), however, the curves and the irregular terrain reduced that perception range, implying shorter distances. The AV was able to identify the obstacle, warn the driver and eventually brake at a fair distance from the obstacle when it was operated on straight roads. This condition corresponded to the AV's high capability. On curvy and dirt roads, the AV was not able to anticipate the obstacles at a reasonable distance, giving drivers less time to react and avoid triggering the emergency brake. These conditions corresponded to the AV's medium and low capabilities.

In the scenario, drivers also had to simultaneously perform a visually demanding NDRT, consisting of a visual search on a separate touchscreen device that exchanged information with the AV. They performed the NDRT only when the self-driving capabilities were engaged. The behavioral measures taken from the drivers were the same from Chapter IV: their focus on the NDRT (from an eye tracker); their ADS usage rate; and their NDRT performance, measured by the number of correctly performed visual searches per second. Drivers were penalized if the emergency brake was triggered, which gave them a sense of the costs and risks of neglecting the AV operation. Specific details about the tasks are given in Section 5.3.3.

The block diagram in Figure 5.3 presents our proposed trust management framework, composed of two main blocks: the trust estimator and the trust calibrator.

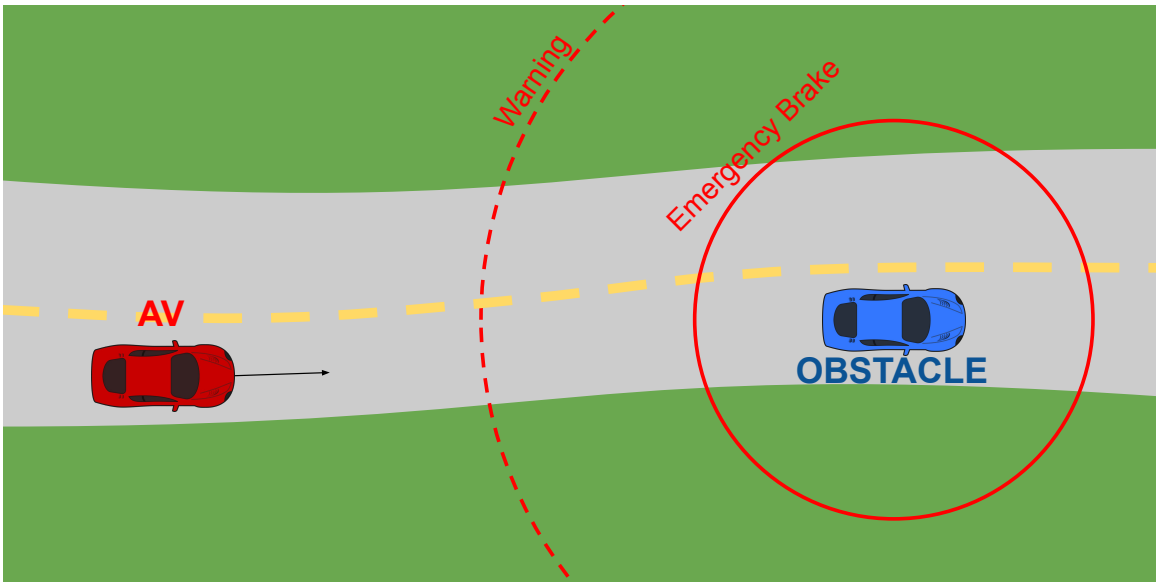

Figure 5.2: Concentric circles represent the distances for which the warning message "Stopped vehicle ahead!" was provided to the driver, and the emergency brake was triggered. The distances varied according to the difficulty of the road. If the emergency brake was triggered, the drivers were penalized on their NDRT score.

The AV block represents elements of the vehicle, such as the sensors to monitor the environment and the ability to output verbal messages to interact with the driver. We present the definitions and the notation used in this chapter in Table 5.1.

5.2.2 Trust Estimator

Figure 5.3 illustrates the trust estimator block, with the AV's alarms and the observation variables φ_k, υ_k and π_k as inputs, and a numerical estimate of drivers' trust in the AV as the output T_k. The observation variables capture the drivers' behavior, which is affected by drivers' trust in the AV. This trust estimator is a simplified version of what is presented in Chapter IV and in [5], and was chosen because of its simple implementation and proven ability to track drivers' trust. Alternative trust estimators could be integrated to the proposed trust management framework if the inputs they require can be captured in real-time. Differently from Chapter IV, we considered that the alarms L_k were always reliable (true alarms), and could not be

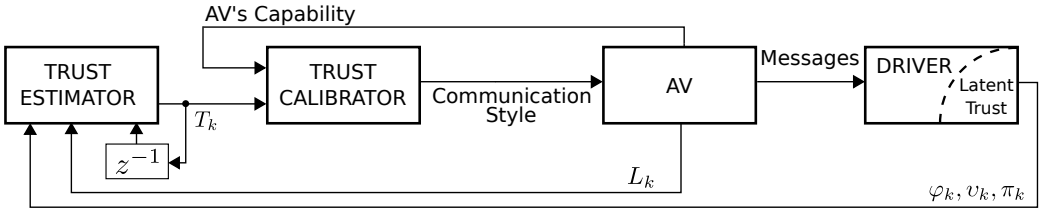

Figure 5.3: Block diagram that represents the trust management framework. The trust estimator block provides a trust estimate T_k to the trust calibrator, which compares it to the capabilities of the AV during operation. The calibrator then defines the communication style that the AV should adopt, and the AV provides the corresponding verbal messages to the driver. L_k represents an alarm provided by the ADS when an obstacle on the road is identified. The observation variables φ_k, υ_k and π_k represent drivers' behaviors, from which drivers' "real" trust (considered a latent variable) is estimated. A delay of one event is represented by the z^{-1} block.

Table 5.1: Definitions and notation used in this chapter

Definition, notation	Characterization
Trial, $[t_0, t_f] \in \mathbb{R}^+$	Trials occur when drivers operate the vehicle on a predefined route, and are characterized by their corresponding time intervals.
Events, $k \in \mathbb{N} \setminus \{0\}$	Events occur each time the ADS warns the driver about an obstacle on the road at t_k, $t_0 < t_k < t_f$.
Alarm, $L_k \in \{0, 1\}$	Boolean variable that is set when the AV correctly identifies an obstacle and warns the driver at the event k. It is reset after the driver passes the obstacle
Focus, $\varphi_k \in [0, 1]$	Drivers' focus on the NDRT, the ratio of time the driver spends looking at the NDRT screen during $[t_k, t_{k+1})$.
Usage, $\upsilon_k \in [0, 1]$	Drivers' ADS usage, the ratio of time the driver spends using the AV's self-driving capabilities during $[t_k, t_{k+1})$.
Performance, $\pi_k \in \mathbb{R}$	Drivers' NDRT performance, the number of points obtained on the NDRT during $[t_k, t_{k+1})$, divided by $\Delta t_k = t_{k+1} - t_k$.
Trust in the AV, $T_k \in [0, 100]$	Drivers' estimated trust in the AV. It is assigned to the interval $[t_k, t_{k+1})$, computed from φ_k, υ_k, π_k and is associated with the covariance Σ_T.

false alarms or misses.

The discrete LTI state-space model for trust dynamics has the form of Eq. (5.1),

$$T_{k+1} = \mathbf{A}T_k + \mathbf{B}L_k + u_k, \tag{5.1a}$$

$$\begin{bmatrix} \varphi_k \\ \upsilon_k \\ \pi_k \end{bmatrix} = \mathbf{C}T_k + \mathbf{w}_k. \tag{5.1b}$$

T_{k+1}, the trust estimate at the event $k+1$, depends on T_k, the alarm L_k, and the process noise u_k. The observation variables depend on the estimated trust and output noise $\mathbf{w}_k$. $\mathbf{A} = \begin{bmatrix} 1.0 \end{bmatrix}$; $\mathbf{B} = \begin{bmatrix} 0.40 \end{bmatrix}$; $\mathbf{C} = 10^{-3} \times \begin{bmatrix} 7.0 & 4.2 & 9.2 \end{bmatrix}^\top$; $u_k \sim \mathcal{N}(0, 0.25^2)$; and $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma_\varphi^2, \sigma_\upsilon^2, \sigma_\pi^2))$, with $\sigma_\varphi = 1.8 \times 10^{-4}$, $\sigma_\upsilon = 7.0 \times 10^{-5}$ and $\sigma_\pi = 5.7 \times 10^{-2}$. (Please see Table 5.1 for variables' definitions.) The parameters for Eq. (5.1) are found by fitting linear models [106] using a previously obtained data set. The state-space structure permits the application of Kalman filter-based techniques for the estimator design. The trust estimator is initialized with

$$T_0 = \frac{1}{3}\left(\frac{\varphi_0}{c_1} + \frac{\upsilon_0}{c_2} + \frac{\pi_0}{c_3}\right), \tag{5.2}$$

where φ_0, υ_0 and π_0 measured over the interval $[t_0, t_1)$ and c_1, c_2, c_3 are the entries of $\mathbf{C}$.

5.2.3 Trust Calibration

The trust calibrator block represented in Figure 5.3 was intended to affect drivers' situation awareness (or risk perception) by changing the communication style of the AV, with the goal of influencing drivers' trust in the AV [77]. At every event k, the AV interacted with the driver through verbal messages corresponding to the communication style defined in the trust calibrator block. The AV can encourage the driver

to focus on the NDRT, moderately warn the driver about the difficulties of the road ahead, or harshly warn the driver, literally demanding driver's attention. Table 5.2 presents the messages the AV provided to the driver in four different communication styles.

To identify trust miscalibrations, the trust calibrator compares the trust estimates with the capability of the AV. Lee and See [60] considered both trust in the automated system and the capabilities of the system as continua that must be comparable to each other. We assumed that the AV's capability corresponds to the three difficulty levels of the road where the AV is operated. We divided the interval $[0, 100]$, for which drivers' trust in the AV was defined, into three sub-intervals: $[0, 25)$ corresponding to low trust, $[25, 75)$ corresponding to medium trust and $[75, 100]$ corresponding to high trust. The uneven distribution of the sub-interval lengths was chosen to mitigate the uncertainty involved in trust estimation. We fit a wider range of values in the medium level, and considered as "low trust" or "high trust" only the estimates that were closer to 0 or 100, respectively. The quantization of both the driver's trust in the AV and the AV's capability in three levels facilitates the real-time comparison of these metrics. Moreover, it permits the definition of a finite set of rules for the trust miscalibration issues. Depending on the application context, alternative quantizations or AV capabilities distributions can be implemented without significant changes to the trust calibrator's framework.

A trust miscalibration is identified whenever there is a mismatch between the AV's capabilities and the driver's level of trust in the AV. The communication style of the AV is then selected after the trust miscalibration is identified. At every event, this comparison results in the identification of one of four distinct driver trust states: undertrusting the AV (*Under*); having an appropriate level of trust in the AV (*Calibrated*); overtrusting the AV (*Over*); or extremely overtrusting the AV (*X-over*). Figure 5.4 shows the ruleset and the correspondence with the resultant communi-

Table 5.2: Messages provided by the AV in each Communication Style

AV Communication Style	Message
Encouraging	"Hey, this is an easy road. You don't need to worry about driving. I will take care of it while you focus on finding the Qs."
Silent	[No message]
Warning (moderate)	"Hey, this part of the road is not very easy. You can still find the Qs, but please pay more attention to the road."
Warning (harsh)	"Look, I told you! I do need your attention. I can feel the road is terrible. I don't know if I can keep us totally safe!"

cation styles of the AV. Note that the establishment of three levels for trust and AV capability is able to cover the occurrence of both undertrust and overtrust, and also allows the identification of extreme overtrust. Extreme overtrust occurs when a driver has a high level of trust in the AV while the AV's capability is low, which is likely to be crucial for driver safety. Therefore, we consider extreme overtrust a trust miscalibration issue that should be seriously addressed.

5.3 Methods

A total of 40 participants ($\mu_{AGE} = 31$; $\sigma_{AGE} = 14$ years) were recruited to take part in the study. From these, 18 were female, 21 male and 1 preferred not to specify gender. We used emails and specialized advertising on the University of Michigan's web portal for behavioral and health studies recruitment. All regulatory ethical concerns were taken, and the study was approved by the University of Michigan's Institutional Review Board.

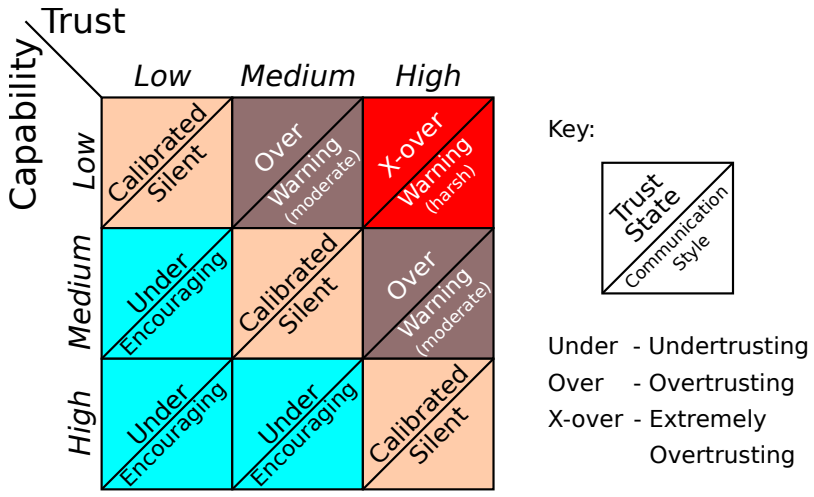

Figure 5.4: Rule set for the trust calibrator. The driver's trust state and the communication style are defined when the AV compares its capability and the driver's trust level. E.g.: when trust is lower than the AV's capabilities (light blue cells), the driver is undertrusting the AV, and the encouraging communication style is selected.

5.3.1 Procedure

Participants signed a consent form and filled out a pre-experiment survey as soon as they arrived at the experiment location. Next, the functions of the AV and the experiment dynamics were explained, and a training drive allowed participants to get familiar with both the AV simulator controls and the NDRT. Participants put the eye-tracker device on and, after it was calibrated, started their first trial on the AV simulator. After the trial, they filled out a post-trial survey. Next, they had their second trial and filled out the post trial survey for the second time. Each experiment took approximately 1 h, and the participants were compensated for taking part in the study. The compensation varied accordingly to their highest total number of points obtained in the NDRT, considering both of their trials. Minimum compensation was of $15, and the participants were able to achieve $20, $30 or $50 in total with a performance cash bonus.

5.3.2 Conditions Randomization

All participants experienced one trial with the trust calibrator and one trial without the trust calibrator. To avoid the participants driving in exactly the same conditions in both of their trials, we varied the direction of the driving on the circuit track. Participants drove in clockwise direction (i.e., regular direction) and counter-clockwise direction (i.e., reverse direction), as mentioned in Figure 5.1. The "trust calibrator use" $\times$ "drive direction" conditions were randomly assigned, depending on the participant's sequential identification number.

5.3.3 Tasks and Apparatus

The driving task was implemented with AirSim over Unreal Engine [108]. The visual search NDRT was the same from Chapter III and Chapter IV, consisting of finding "Q" characters among a field of "O" characters. The NDRT was implemented with PEBL [80]. Participants' scores increased by 1 point every time they correctly selected the targets on the screen, and they lost 20 points each time the emergency brake was activated. Source codes for both tasks are available at `https://github.com/hazevedosa/tiavManager`. The experimental setup is shown in Figure 1.3.

5.4 Results

We analyzed the impacts of using the trust calibrator's adaptive communication with different communication styles on drivers' trust in the AV (i.e., real-time estimated trust T_k). For this, we analyzed the differences in drivers' trust estimates between consecutive events after they had heard the messages from the AV. Drivers' trust differences are given by $\Delta T = T_k - T_{k-1}$, i.e., the difference between trust estimates after and before the event k. ΔT was specifically computed for the analysis, and indicates how participants' trust estimates changed after they were encouraged

or warned by the AV at the event k (i.e., after the AV interacted with the drivers adopting the communication style corresponding to drivers' trust states at the event k).

Drivers showed significant positive or negative differences in their trust estimates after the AV encouraged or warned them. Table 5.3 and Figure 5.5 present the results obtained with a linear mixed-effects model for ΔT. Linear mixed-effects models are regression models that include both fixed and random effects of independent variables on a dependent variable. Fixed effects represent the influence of the independent variables or treatments of primary interest (in this case, the communication styles) on the dependent variable (i.e., trust difference ΔT). Random effects represent differences that are not explained by the factors of primary interest but are rather related to hierarchical organizations present in the sample population (e.g., groups of data collected from the same participant) [106]. For instance, in this analysis, a random intercept for each participant in the experiment was added to the ΔT linear mixed-effects model. In summary, we sought the β parameters that best fit the model

$$\Delta T = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_p, \tag{5.3}$$

where $x_1 = 1$ when the communication style was "Encouraging" and $x_1 = 0$ otherwise; $x_2 = 1$ when the communication style was "Warning (moderate)" and $x_2 = 0$ otherwise; and $x_3 = 1$ when the communication style was "Warning (harsh)" and $x_3 = 0$ otherwise. The random effect ϵ_p had mean $\mu = 0$ and standard deviation $\sigma = 25.3$, and represented each participant's characteristic intercept and the irreducible error of the model. Table 5.3 shows that all β parameter estimates corresponding to the non-silent communication styles were significant ($p < 0.01$).

Table 5.3: Communication style fixed effects on drivers' Trust in AV difference (ΔT), obtained with a linear mixed-effects model [106]

Trust State/ Communication Style	Parameter	Estimate	Standard Error (S.E.)	Student's t	p-value	Lower Bound	Upper Bound
Calibrated/ Silent*	β_0	1.7	1.8	0.92	0.36	-1.9	5.3
Under/ Encouraging	β_1	**+15.4**	3.3	4.7	3.3×10^{-6}	9.0	21.8
Over/ Warning (moderate)	β_2	**-9.0**	2.8	-3.2	1.7×10^{-3}	-14.6	-3.4
X-over/ Warning (harsh)	β_3	**-22.9**	5.1	-4.5	9.9×10^{-6}	-33.0	-12.8

Obs.: *Model intercept reference; Significant parameter estimates ($p < 0.01$) in bold font. A random intercept is assigned to each participant in the data set.

In general, the reaction of the drivers to the AV messages followed an expected trend. The lack of messages did not significantly change driver's behaviors when their trust in the AV was calibrated: the average difference—considered the reference intercept for the linear mixed-effects model—was 1.7 units, but the p-value of 0.36 indicates that it was not significantly different from 0. The encouraging messages helped drivers to increase their trust in the AV: as shown in Table 5.3, the average increase was $1.7 + 15.4 = 17.1$ units for undertrusting drivers. The warning messages had the effect of decreasing their trust in the AV: trust estimates of overtrusting drivers varied by $1.7 - 9.0 = -7.3$ units, and for extremely overtrusting drivers, trust estimates varied by $1.7 - 22.9 = -21.2$ units. Figure 5.6 exemplifies the time trace for a participant's trust estimates during a trial, indicating the messages provided by the AV and the regions for which trust would be considered calibrated.

The use of the calibrator reduced trust miscalibrations for 29 (out of 40) participants. We computed *trust miscalibration time ratios*, representing the amount of time drivers' trust state was different from "Calibrated", relative to the total time of each trial. For the computation, we removed the intervals right after a change in AV's capabilities, where miscalibrations were intentionally caused. For all participants, the average trust miscalibration time ratio was 70% in trials for which the calibrator was not used. This ratio was reduced to 43.7% when the calibrator was used. Considering only the 29 participants that had their miscalibration time ratios reduced (when using the trust calibrator), these ratios were 82% and 42%, respectively. For the remaining 11 participants, the reasons for their lack of decreased trust miscalibration are unknown, although we believe these reasons could be related to the limitations imposed by the short duration of the experiment.

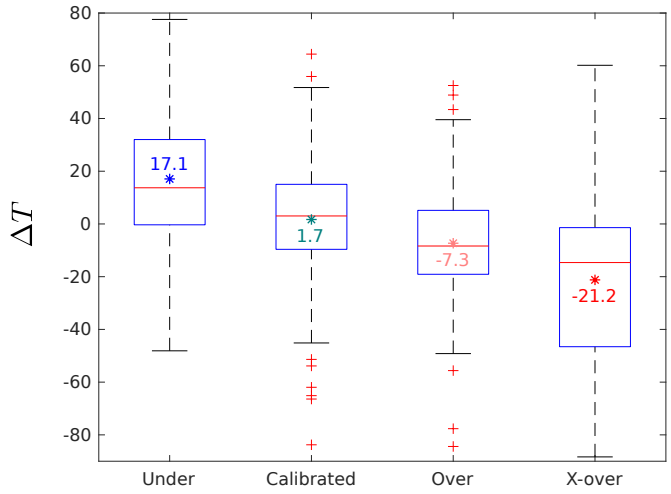

Figure 5.5: Distributions of drivers' trust in the AV differences (ΔT), for the different driver trust states. Overtrusting drivers received the warning communication styles and responded with negative differences. Undertrusting drivers received the encouraging communication styles and responded with positive differences. Drivers with calibrated trust had relatively small positive differences on average. The average values were obtained from the parameter estimates in Table 5.3.

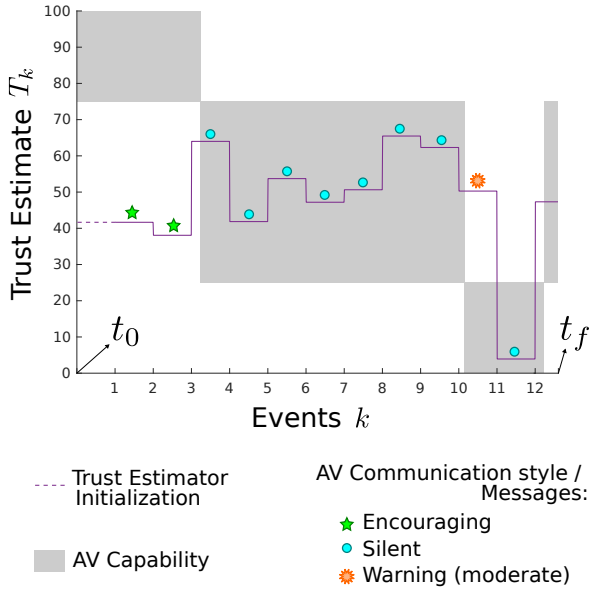

Figure 5.6: Time trace for a driver's trust estimates T_k, which is assigned to the interval $[t_k, t_{k+1})$ after being computed from φ_k, υ_k and π_k. After two encouraging messages when the driver undertrusted the AV, T_k increased. After a warning message when the driver overtrusted the AV, T_k decreased. While driver's trust was calibrated, the calibrator refrained from providing messages to the driver.

5.5 Discussion

The results presented in Section 5.4 support the effectiveness of our trust management framework or, more specifically, our trust calibrator, which is the main intended contribution of this chapter. When undertrusting drivers increase their trust in the AV, their trust state is likely to approach the condition of trust calibration. Equivalently, when overtrusting drivers decrease their trust in the AV, they are more likely to reach trust calibration. The increase of trust for undertrusting drivers means that after the communication from the AV, drivers were able to use the self-driving capabilities more confidently, which was reflected by the increases of their related observation variables. Likewise, the framework was able to reduce drivers' trust levels if they presented overtrusting behaviors, when the driving context was not favorable to the AV's autonomous operation. The AV communication demanding drivers' attention to the driving task was effective, tending to adjust (i.e., decrease) drivers' behaviors when they overtrusted the AV.

The proposed real-time trust calibration method was inspired by the relationships among situation awareness, risk perception and trust. Previous works reported on the effectiveness of situation awareness and perceived risk to impact drivers' trust in AVs [4, 92, 94, 139]. We applied different communications styles and messages in an attempt to vary drivers' situation awareness and risk perception in real time. In consequence, we deliberately induced equivalent real-time changes in trust, supporting the drivers to avoid trust miscalibrations by reducing the difference between their trust estimates and the AV's capability references. The main applicability of the proposed trust management framework is to enable AVs to perceive drivers' trusting behaviors and react to them accordingly. Smart ADSs featuring this capability would likely enhance the collaboration between the driver and the AV because it permits the adaptation of attentional resources according to the operational environment and situation.

Our method can be considered a complement to [1] and [67]. The work in [67] supported our insights for the use of eye-tracking-based techniques for real-time trust estimation. In comparison to [1], we used different methods and behavioral variables for trust estimation and extended their ideas to include the trust calibrator and propose our trust management framework.

The limitations of the management framework are mostly related to the uncertainty involved in influencing drivers' trust with different messages, which might not be very effective for some drivers. These drivers might need several interactions to be persuaded by the AV. An example is illustrated in Figure 5.6, where the driver was encouraged to trust the AV twice before the increase in $\Delta T = T_3 - T_2$ was registered. The spreads of the box plots represented in Figure 5.5 suggest that, in less frequent cases, drivers could present an unexpected behavior, not complying with AV's encouraging or warning messages. The lack of a process for customizing the parameters of our framework contributes to this uncertainty. Relying on average model parameters in the trust estimation block can reduce the accuracy of the estimates because the parameters of each driver can be very different from the averages. Therefore, the trust estimation algorithm (and consequently, the management framework) might work more efficiently if adapted to each individual driver. Another limitation is that the capability of the AV was defined by the circuit track difficulty levels only. Other factors can affect AV capability and could be considered, such as those related to vehicular subsystems or to the weather.

5.6 Conclusions and Contribution

The main contribution of this chapter is the proposed trust calibration method, which is used in the framework for managing drivers' trust in AVs in order to avoid trust miscalibration issues. The framework relies on observing drivers' behaviors to estimate their trust levels, comparing it to capabilities of the AVs, and activating dif-

ferent communication styles to encourage undertrusting drivers and warn overtrusting drivers. Our proposed management framework has shown to be effective in inducing positive or negative changes on drivers' trust in the AV and, consequently, mitigating trust miscalibration.

The proposed trust management framework is applicable to intelligent driving automation systems, providing them with the ability to perceive and react to drivers' trusting behaviors, improving their interaction with the AVs, and maximizing their safety and their performance in tasks other than driving. However, this framework can not assess whether the driver can be trusted or not to take over control of the vehicle when necessary. Chapter VI advances in this direction and proposes a bi-directional trust model that could be used for modeling both human and robotic trust.

CHAPTER VI

Bi-Directional Model for Natural and Artificial Trust

6.1 Introduction

Unlike traditional automation, autonomous robots could adjust their behaviors depending on how their human counterparts appear to be trusting them or how humans appear to be trustworthy. This chapter introduces a novel capabilities-based bi-directional multi-task trust model that can be used for trust prediction either from a human or from a robotic trustor agent. Tasks are represented in terms of their capability requirements, while trustee agents are characterized by their individual capabilities. Trustee agent's capabilities are not deterministic; rather, they are represented by belief distributions. For each task to be executed, a higher level of trust is assigned to trustee agents who have demonstrated that their capabilities exceed the task's requirements. Results of an online experiment with 284 participants are reported, and reveal that the proposed model outperforms existing models for multi-task trust prediction from a human trustor. Simulations of the model for determining trust from a robotic trustor are also presented. This bi-directional trust model is intended to be useful for applications involving control authority allocation in human-robot teams.

This chapter is based on the research directions published in [8] and the work published in [9]. The remainder of this chapter is organized as follows. Section 6.2 describes the development of a bi-directional trust model that can be used in situations where a human collaborates with a robot. This model can be used for representing both the human's "natural" trust and the robot's "artificial" trust. Section 6.3 presents an online experiment that was conducted to obtain the data for validating the proposed bi-directional trust model. Section 6.4 focuses on the results obtained, both for the prediction of human-drivers' natural trust in robotic AVs and for the definition of a robot's artificial trust in humans. Section 6.5 discusses the main strengths and limitations of the proposed bi-directional trust model and Section 6.6 concludes the chapter.

6.2 Bi-Directional Trust Model Development

6.2.1 Context Description

Consider the following situation: two agents (human H or robot R) collaborate and must execute a sequence of tasks. These tasks are indivisible, and must be executed by only one agent. The execution of each task can either succeed or fail. For each task, one of the agents will be in the position of trustor, and the other will be the trustee. Therefore, the trustor will be vulnerable to the trustee's performance in that task. From previous experiences with the trustee, the trustor has some implicit knowledge about the trustee's capabilities. This implicit knowledge is used by the trustor assess how likely is the trustee to succeed or fail in the execution of a task.

6.2.2 Definitions

We define the terms and concepts we need for developing our trust model:

Definition 1 - Task. A *task* that must be executed is represented by $\gamma \in \Gamma$. Γ

represents the set of all tasks that can be executed by the agents.

Definition 2 - Agent. An *agent* $a \in \{H, R\}$ represents a trustee that could execute a task γ.

Definition 3 - Capability. The representation of a specific skill that agents have/that are required for the execution of tasks. We represent a capability as an element of a closed interval $\Lambda_i = [0, 1]$, $i \in \{1, 2, 3, ..., n\}$, with n being a finite number of dimensions characterizing distinct capabilities.

Definition 4 - Capability Hypercube. The compact set representation of n distinct capabilities, given by the Cartesian product $\Lambda = \prod_{i=1}^{n} \Lambda_i = [0, 1]^n$. This definition is inspired by the particular capabilities from Mayer's model [69], namely ability, benevolence and integrity, but the definition is intended to be broader than these three dimensions.

Definition 5 - Agent's Capability Transform. The agent capability transform $\xi : \{H, R\} \to \Lambda$ maps an agent into a point in the capability hypercube representing the *agent's capabilities*, given by $\xi(a) = \lambda = (\lambda_1, \lambda_2, ..., \lambda_n) \in \Lambda$.

Definition 6 - Task Requirements Transform. The *task requirements transform* $\varrho : \Gamma \to \Lambda$ maps a task γ into the minimum required capabilities for the execution of γ, given by $\varrho(\gamma) = \bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2, ..., \bar{\lambda}_n) \in \Lambda$.

Definition 7 - Time Index. The *time* is discrete and represented by $t \in \mathbb{N}$.

Definition 8 - Task Outcome. The outcome of a task γ after being executed by the agent a at the time t is represented by $\Omega(\xi(a), \varrho(\gamma), t) \in \{0, 1\}$, where 0 represents a failure and 1 represents a success. We also define the Boolean complement of Ω, denoted by $\mho$, therefore being $\mho = 1$ when $\Omega = 0$, and $\mho = 0$ when $\Omega = 1$.

Leveraging the previous definitions, trust can be finally defined.

Definition 9 - Trust. A trustor agent's trust in a trustee agent a to execute a

task γ at a time instance t can be represented by

$$\begin{aligned}\tau(a,\gamma,t) &= P\big(\Omega(\xi(a),\varrho(\gamma),t)=1\big)\\ &= \int_{\Lambda} p\big(\Omega(\lambda,\bar{\lambda},t)=1|\lambda,t\big)bel(\lambda,t-1)d\lambda,\end{aligned} \tag{6.1}$$

where $\lambda = \xi(a)$, $\bar{\lambda} = \varrho(\gamma)$, and $bel(\lambda, t-1)$ represents the trustor's belief in the agent's capabilities λ at time $t-1$ (i.e., before the actual task execution). The belief is a dynamic probability distribution over the capability hypercube Λ. Note that, at each time instance t, trust is a function of the task requirements $\bar{\lambda}$, representing a *probability of success* in $[0,1]$. This formulation is consistent with the definition presented in [115].

6.2.3 Bi-directional Trust Model

Our bi-directional model is defined by Eq. (6.1), and depends on the combination of:

- a function to represent the "trust given the trustee's capability", represented by the conditional probability $p\big(\Omega(\lambda,\bar{\lambda},t)=1|\lambda,t\big)$; and
- a process to dynamically update the trustor's belief over the trustee capabilities $bel(\lambda,t)$.

We assume that an agent that successfully performs a task is more likely to be successful on less demanding tasks. Conversely, an agent that fails on a task is more likely to fail on more demanding tasks. We adapt the sigmoid function to represent that logic, and for each capability dimension we can write

$$\tau_i = \left[\frac{1}{1+e^{\beta_i(\bar{\lambda}_i-\lambda_i)}}\right]^{\zeta_i}, \tag{6.2}$$

where $\beta_i, \zeta_i > 0$. We call the β_i parameter the trustor's *pragmatism*, while the ζ_i

parameter is called the trustor's *skepticism*, both for the i-th capability dimension. Considering that all capability dimensions must be assessed concurrently and assuming that the capability dimensions are represented by independent random variables, for the probability computation, we have

$$p\big(\Omega(\lambda, \bar{\lambda}) = 1|\lambda\big) = \prod_{i=1}^{n} \tau_i = \prod_{i=1}^{n} \left[\frac{1}{1 + e^{\beta_i(\bar{\lambda}_i - \lambda_i)}}\right]^{\zeta_i}, \tag{6.3}$$

where t was suppressed, as the resulting function is independent of the time. The product of probabilities in Eq. (6.3) can quickly converge to zero as n increases. Therefore, to improve code implementation stability in practical implementations, a linear form of Eq. (6.3) could be used (i.e., by taking the logarithm on both sides of the equation).

Trust dynamics is established with a process for updating $bel(\lambda, t)$ that relates observations of a trustee agent's past performances with that agent's likelihood of success on related tasks. We consider that a trustor agent must build the belief about the trustee's capabilities after observations of the trustee's performances. However, initially, the trustor has no information about the trustee's performances and capabilities. We assume this is represented by $bel(\lambda, 0)$ being a uniform probability distribution over the capability hypercube Λ, i.e., $bel(\lambda_i, 0) = \mathcal{U}(0, 1), \forall i \in \{1, 2, ..., n\}$. Next, after observing the sequence of successes and failures of the trustee in different tasks, the trustor updates $bel(\lambda, t)$, following the procedures in Algorithm 2 and in Figure 6.1

6.2.4 Artificial Trust

For representing the artificial trust of a robotic trustor in a trustee agent, the bi-directional trust model can be slightly modified. We can vanish the subjective biases that characterize human trustors by considering large values for the parameters

Algorithm 2 Capability Belief Initialization and Update

procedure CAPABILITY HYPERCUBE INITIALIZATION
  for $i = 1 : n$ **do**
    $\ell_i \leftarrow 0$
    $u_i \leftarrow 1$
    $bel(\lambda_i, 0) \leftarrow \mathcal{U}(\ell_i, u_i)$ ▷ *Uniform distributions*
  end for
end procedure
procedure CAPABILITY UPDATE($\gamma, bel(\lambda, t-1)$)
  ▷ *When trustor observes trustee executing* γ *at* t
  for $i = 1 : n$ **do**
    if $\Omega(\lambda, \bar{\lambda}, t) = 1$ **then**
      if $\bar{\lambda}_i > u_i$ **then**
        $u_i \leftarrow \bar{\lambda}_i$
      else if $\bar{\lambda}_i > \ell_i$ **then**
        $\ell_i \leftarrow \bar{\lambda}_i$
      end if
    else if $\Omega(\lambda, \bar{\lambda}, t) = 0$ **then**
      if $\bar{\lambda}_i < \ell_i$ **then**
        $\ell_i \leftarrow \bar{\lambda}_i$
      else if $\bar{\lambda}_i < u_i$ **then**
        $u_i \leftarrow \bar{\lambda}_i$
      end if
    end if
    $bel(\lambda_i, t) \leftarrow \mathcal{U}(\ell_i, u_i)$
  end for
end procedure

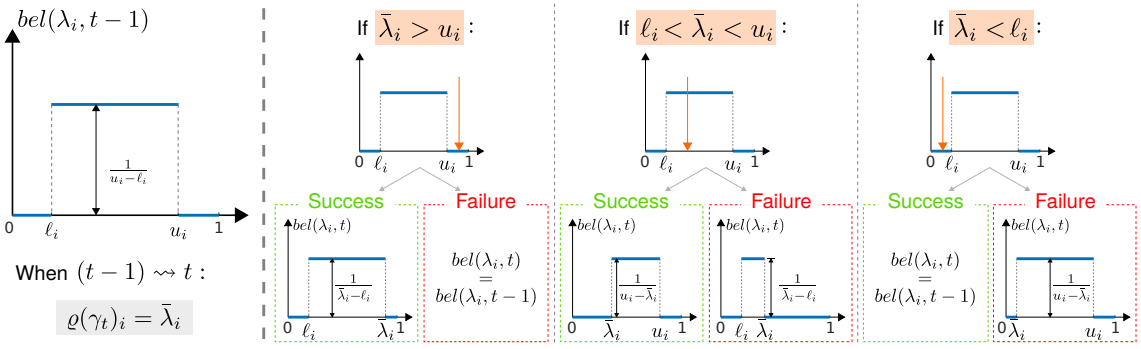

Figure 6.1: Capability update procedure, where each capability dimension changes after the trustor agent observes the trustee agent a executing a task γ_t (at a specific time instance t). The belief distribution over a's capabilities *before* the task execution $bel(\lambda_i, t-1)$ is updated to $bel(\lambda_i, t)$, depending on the task capability requirements $\varrho(\gamma_t)_i = \bar{\lambda}_i$ and on the performance of a in γ_t, which can be a success ($\Omega = 1$) or a failure ($\Omega = 1$).

β_i in Eq. (6.2). It makes the robot *infinitely pragmatic*, and its trust given the trustee agent's capability is reduced to 1 for all tasks with requirements less than that capability and to 0 for all tasks with requirements greater than that capability. We achieve this by using a sufficiently large value for β_i, for which τ_i becomes an analytic approximation of a decreasing step function with the transition from 1 to 0 when $\bar{\lambda}_i = \lambda_i$, i.e.

$$\lim_{\beta_i \to \infty} \tau_i = \mathcal{H}\big(- \bar{\lambda}_i + \lambda_i \big), \tag{6.4}$$

where $\mathcal{H}(x)$ is the Heaviside function of a continuous real variable x. Considering all capability dimensions to be independent, and using the approximation in Eq. (6.4) for computing trust with Eq. (6.3) and Eq. (6.1), we have

$$\tau(a, \gamma, t) = \prod_{i=1}^{n} \psi(\bar{\lambda}_i), \tag{6.5}$$

where,

$$\psi(\bar{\lambda}_i) = \begin{cases} 1 & \text{if } 0 \leq \bar{\lambda}_i \leq \ell_i, \\ \frac{u_i - \bar{\lambda}_i}{u_i - \ell_i} & \text{if } \ell_i < \bar{\lambda}_i < u_i, \\ 0 & \text{if } u_i \leq \bar{\lambda}_i \leq 1. \end{cases} \tag{6.6}$$

Therefore, for each capability dimension, the robotic trustor agent believes that the trustee agent's capability is a random variable λ_i uniformly distributed between ℓ_i and u_i. If a task requires $\bar{\lambda}_i < \ell_i$, the trustee capability exceeds the task requirement and trust is 1. Conversely, if $\bar{\lambda}_i > u_i$, the task requirement exceeds the trustee's capability and trust is 0. In the intermediate condition, trust decreases with a constant slope from 1 to 0, corresponding to $\bar{\lambda}_i = \ell_i$ and $\bar{\lambda}_i = u_i$ respectively.

Differently from humans, robots can store accurate information for a long time, and can use this long-term information to update their capability beliefs with a process different from that presented in Algorithm 2. An alternative is to recursively solve an optimization problem, considering the history of outcomes observed from different

tasks γ (with different $\varrho(\gamma) = \bar{\lambda} \in \Lambda$). Trust is approximated by the number of successes divided by the number of times the task γ was performed, i.e.,

$$\hat{\tau} = \frac{\sum_{m=0}^{t} \Omega(\xi(a), \varrho(\gamma), m)}{\sum_{m=0}^{t} \left[\Omega(\xi(a), \varrho(\gamma), m) + \mho(\xi(a), \varrho(\gamma), m)\right]}, \tag{6.7}$$

and, considering each $\lambda = \varrho(\gamma)$, the capability distribution limits ℓ_i and u_i should be chosen such that $bel(\lambda, t) = \prod_{i=1}^{n} \mathcal{U}(\hat{\ell}_i, \hat{u}_i)$, and

$$(\hat{\ell}_i, \hat{u}_i) = \underset{[0,1]^2}{\arg\min} \int_{\Lambda} \|\tau - \hat{\tau}\|^2 d\lambda. \tag{6.8}$$

For numerical computations, Λ can be discretized and Eq. (6.8) approximated with a summation, as in Section 6.4.2.

6.3 Experiment

We conducted an online experiment using a Qualtrics survey and the Amazon Mechanical Turk (MTurk) platform to gather a dataset for comparing our model with other trust prediction models, such as Soh's models [115] and OPTIMo [132]. We aimed to emulate a human-AV interaction setting, asking participants to (1) assess the requirement levels for driving tasks that were to be executed by the AV; (2) watch videos of the AV executing a part of those tasks; and (3) evaluate their trust in the AV to execute other tasks (distinct from those they have watched in the videos).

Initially, only images and verbal descriptions of four driving tasks were presented in random order to the participants. These images and descriptions are presented in Figure 6.2. Participants were asked to rate the capability requirements for each of the presented tasks in terms of two distinct capabilities of the AV: sensing and processing. These capability dimensions were defined and presented to the participants as,

- **Sensing (λ_s) -** *The accuracy and precision of the sensors used to map the environment where the AV is located and perceive elements within that environment, such as other vehicles, people, and traffic signs.*

- **Processing (λ_p) -** *The speed and performance of the AV's computers that use the information from sensors to calculate the trajectories and the steering, acceleration, and braking needed to execute those trajectories.*

Participants were asked to indicate the required capability levels $(\bar{\lambda}_s, \bar{\lambda}_p) \in [0, 1]^2$ for each task providing a score (i.e., indicating a slider position on a continuous scale) varying from low to high. Although the limits from 0 to 1 were not directly shown to the participants on the continuous scale, we used the relative positions of the slider markers to compute their answers for the required capability levels.

After evaluating all four presented tasks, participants watched short videos (approximately $20s$ to $30s$) of a simulated AV executing three of the four tasks. Those three were considered *observation tasks*. The videos showed the AV succeeding or failing to execute each observation task. Next, participants were asked to indicate whether the AV did successfully execute the task or not. That question served both as an attention checker and as a way to make the participant acknowledge the performance of the AV in that specific task. After watching each video, participants were also asked to rate their trust τ in the AV to execute the fourth remaining task (i.e., the *trust prediction task*) on a 7-point Likert scale varying from "Very Low Trust" to "Very High Trust", as an indication of how much they disagreed or agreed with the sentence: "*I believe that the AV would successfully execute the task*". They were asked to consider all videos they had seen during the observation tasks and rate their trust in the AV to execute the trust prediction task. Finally, participants received a random 4-digit identifier code to upload in the MTurk platform and receive their payment.

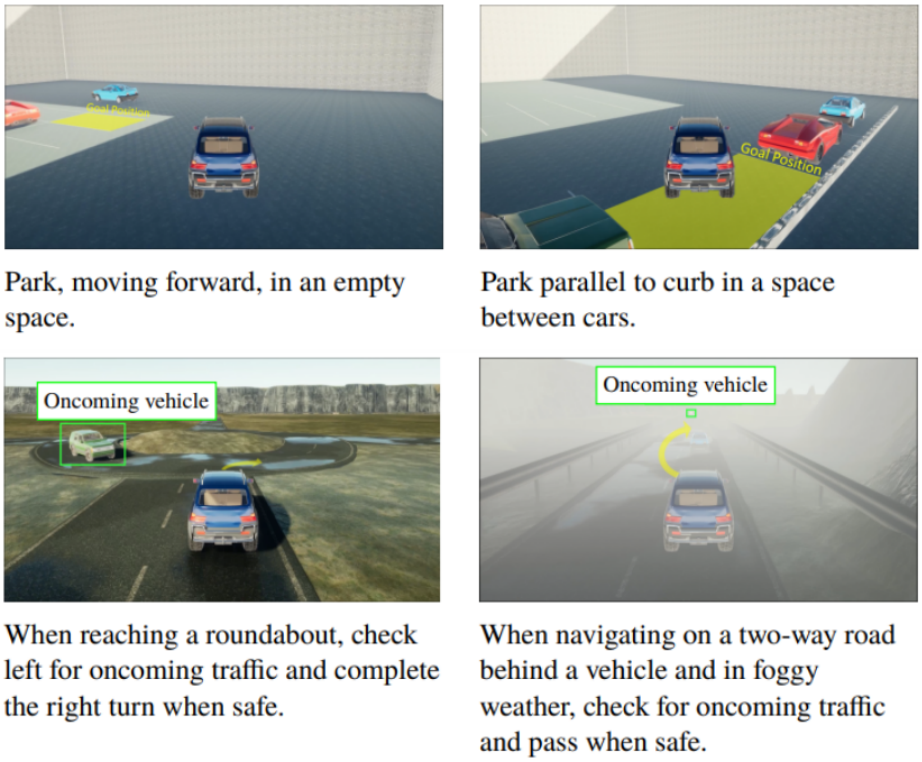

Figure 6.2: Tasks presented to the experiment participants in terms of images and corresponding verbal descriptions. The participants had to rate the capability requirements for each of these tasks, considering two capability dimensions: sensing and processing. In other words, they had to assign a pair $(\bar{\lambda}_1, \bar{\lambda}_2) \in [0, 1]^2$ for each task. Tasks were randomly presented for avoiding ordering effects.

To keep work-related regulations consistent, we restricted our participants to physically be in the USA when accepting the MTurk human intelligence task (HIT). A total of 284 MTurk workers participated in our experiment and received a payment of $1.80 for completing the HIT without failing to correctly answer the attention checker questions. The HITs were completed in approximately 6min40s, on average. We collected no demographics data or other personal information from the participants, as these were not used in our analyses. The obtained dataset and our implementations are available at `https://bit.ly/3sfVtuK`. The research was reviewed and approved by the University of Michigan's Institutional Review Board.

6.4 Results

6.4.1 Human-drivers' (natural) trust in robotic AVs

We implemented a 10-fold cross-validation scheme to train and evaluate our bi-directional trust model (BTM) with the data obtained in the experiment described in Section 6.3. For comparison, we also evaluated the performance of Soh's Bayesian Gaussian Process model (GP) [115] and of a linear Gaussian model similar to Xu and Dudek's OPTIMo (OPT) [132] on our collected dataset. We obtained the tasks' vector representations required for the GP model with GloVe [91], by processing the verbal descriptions presented in Figure 6.2. There were no closed analytical forms for Eq. (6.1), therefore we discretized each task capability dimensions in 10 equal parts and computed numerical approximations for τ. Since we considered only two outcome possibilities (fail or success in executing a task), the trust measurements from both the dataset and the model outputs were considered probability parameters of Bernoulli distributions. We considered the cross-entropy between those distributions to be the loss function to be minimized. We used PyTorch [90] to implement all parameter optimizations with the Adam algorithm [50], using randomized validation sets comprising 15% of the training data. Two metric scores were computed for the comparisons among model performances: the Mean Absolute Error (MAE); and the Negative Log-Likelihood (NLL), which corresponds to the loss function chosen for the optimizations.

Table 6.1 presents the MAE and NLL scores averaged over the 10 cross-validation folds (with standard deviations between parentheses) for the BTM, GP and OPT models. Figure 6.3 complements the table, showing the average learning curves for both scores and bars representing the average final values with ± 1 standard deviations.

Our bi-directional trust model (BTM) outperforms both the GP and the OPT

Table 6.1: Mean Absolute Error (MAE) and Negative Log-Likelihood (NLL) average minimized scores for each trust model

Model	**MAE**†	**NLL**†
BTM	$\mathbf{0.196(0.020)}^{\ddagger}$	$\mathbf{0.593(0.033)}^{\ddagger}$
GP	0.220(0.028)	0.619(0.060)
OPT	0.280(0.016)	0.672(0.021)

†10-fold results: Mean(Standard Deviation).
‡Best scores in **bold**.

models after the parameter optimization process. BTM reduces the MAE metric by approximately 11% as compared with GP, and by 30% as compared to OPT. In terms of NLL, the use of BTM reduces this metric by approximately 4.3% as compared with GP model, and by 12% as compared with the OPT model.

6.4.2 Robots' Artificial Trust in Humans

Besides evaluating and comparing our bi-directional trust model with other trust models using experimental data, we also implemented simulations to verify its use in the artificial trust mode (i.e., as a model for predicting a robots' trust in another trustee agent). To the best of our knowledge, this is the first artificial trust model, and therefore its performance could not be compared to other models such as what was done in Subsection 6.4.1. We assumed two unspecified capability dimensions, considering that a trustee agent a's capabilities were static and represented by a point $\xi(a) = (\lambda_1, \lambda_2) \in \Lambda = [0,1]^2$. The trustee agent's capabilities were initially unknown by the trustor robot, who must estimate $\xi(a)$ after observing the trustee's performances in several different tasks. We considered N fictitious tasks γ^j, $j \in \{1, 2, ..., N\}$, and randomly picked N points $\varrho(\gamma^j) = (\bar{\lambda}_1^j, \bar{\lambda}_2^j) \in \Lambda$ representing capability requirements for the tasks. Task outcomes were assigned to each of the N tasks, with high probability of success for tasks that simultaneously had $\bar{\lambda}_1^j \leq \lambda_1$ and $\bar{\lambda}_2^j \leq \lambda_2$, and low probability of success when $\bar{\lambda}_1^j > \lambda_1$ or $\bar{\lambda}_2^j > \lambda_2$. Again, for

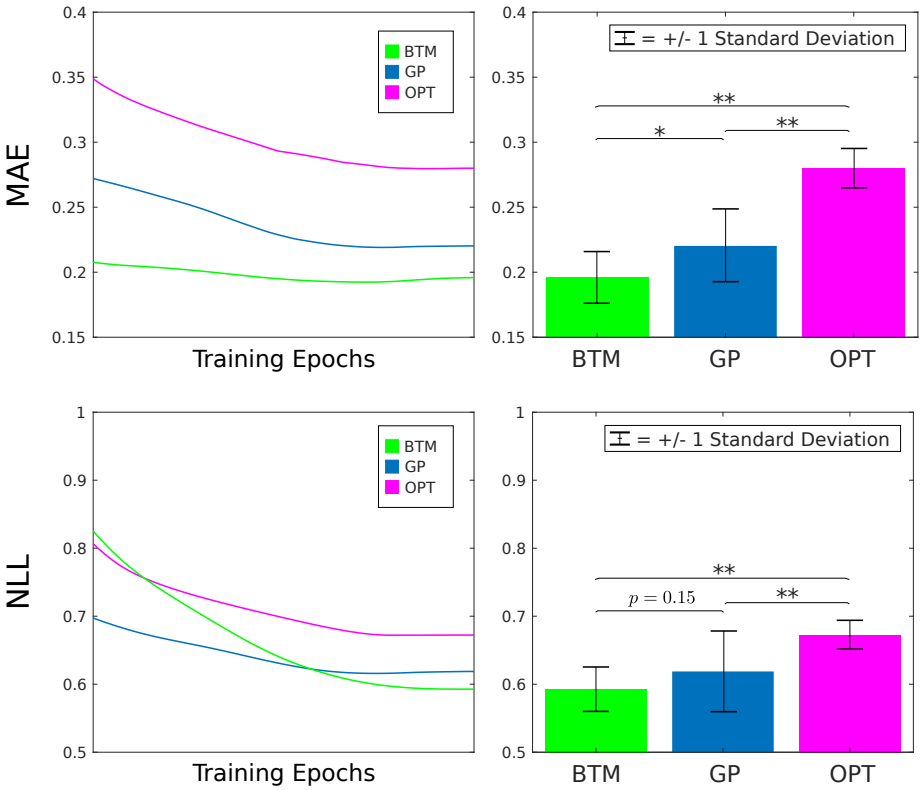

Figure 6.3: MAE and NLL learning curves and final values for our proposed trust model (BTM) and for current trust models from [115] (GP) and [132] (OPT). As the total number of training epochs is different for each model, their representation on the horizontal axes of the learning curves is normalized.*$p < 0.05$; **$p < 0.01$.

numerical computations, we discretized both capability dimensions in 10 equal parts, obtaining 100 bins for Λ. We computed the observed probabilities of success for tasks inside a bin dividing the number of successes by the total number of tasks that fell on each bin (i.e., the approximation for $\hat{\tau}$). Finally we ran optimizations to find the parameters that best characterized $bel(\lambda_1, N)$ and $bel(\lambda_2, N)$, solving the problem represented by Eq. (6.8).

Figure 6.4 illustrates the evolution of $bel(\lambda, N)$ and of $\tau(a, \gamma, N)$ for increasing values of N. The higher the number of observations, the better the accuracy of a's identified capabilities, represented with a smaller gray area in the $bel(\lambda, N)$ distribution.

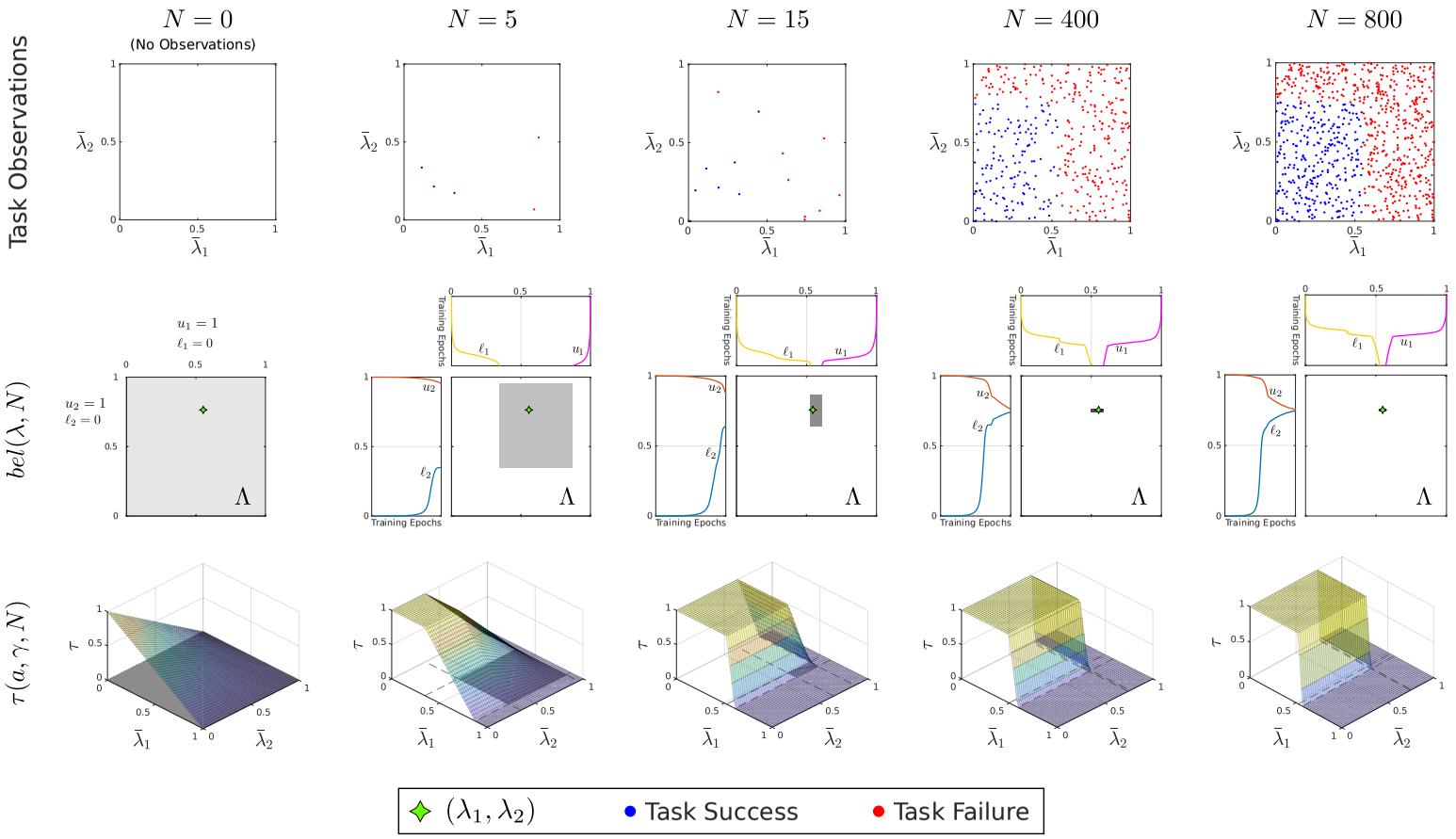

Figure 6.4: Artificial trust results, where a robotic trustor agent's belief over a trustee agent a's capabilities is updated after N observations of a's performances in different tasks, represented by points in $\Lambda = [0, 1]^2$. When $N = 0$, $bel(\lambda, N)$ is "spread" over the entire Λ. When the robot trustor collects observations, it starts building a's capabilities profile and reducing the gray area in the $bel(\lambda, N)$ distribution. This profile gets more accurate when N increases and (λ_1, λ_2) gets better defined. This is also reflected in the evolution of the conditional trust function $\tau(a, \gamma, N)$.

6.5 Discussion

Our model is based on general capability representations that can be either performance or non-performance trust factors. This particular aspect of our bi-directional trust model makes it useful for representing a robot's artificial trust, as presented in Subsection 6.4.2, and allows for better human trust predictions in comparison to existing models, as presented in Subsection 6.4.1. Additionally, our model considers task capability requirements in its description, describing how hard a task is for an agent to execute. The model's mathematical formulation captures the differences between those task requirements and the potential trustee agent's observed capabilities. Differently from the Gaussian process-based method presented in [115], this formulation allows for the adequate representation of lower trust levels when the requirements of a task exceed the capabilities of the agent and, conversely, higher trust levels when the agent capabilities exceed the task requirements.

The results shown in Section 6.4.1 reveal that our proposed bi-directional trust model has better performance for predicting a human's trust in a robot—in our specific experiment, an AV—than the models from [132] and [115]. This performance improvement was expected because current models are limited in capturing important trust-related parameters, such as the agents' capabilities or task's requirements in their formulation. To the best of our knowledge, only our model and Soh's models [115] distinguish and describe the trust transfer between different tasks, while OPTIMo [132] is more appropriate for predicting a human's trust in a robot to execute one specific task.

Section 6.4.2 presents simulations that show how the proposed model can be used for representing a robot's artificial trust. In the future, the proposed bi-directional trust model could be used in real-world human subjects experiments. An example could be a study where participants would execute some tasks represented in the capability hypercube, and the robot would be able to establish its trust in the par-

ticipants based on their failures or successes on those tasks. In parallel, the robot could estimate the human's natural trust for different tasks, and use both natural and artificial trust metrics to compute expected rewards for the execution of new tasks. Tasks could be allocated between the human and the robot to maximize the expected reward of a whole set of tasks, eventually improving the joint performance of the human–robot team.

The dynamic allocation of tasks is likely to require the computation of an agent's *self-trust* in parallel with the computations of trust in the counterpart agent. Although the results of the experiment and of the simulations presented in Section 6.4 have not included the computation of natural or artificial self-trust, we consider that our model can be used for those computations. Our assumptions do not require that the trustor agent and the trustee agent must be distinct and, our best judgment is that there are no reasons to impose this restriction to the trust model applicability. To facilitate (and possibly improve) the prediction of self-trust and trust in the other agent, different strategies for updating capabilities and trust over time can be implemented in parallel with the feedback represented by performance observations. Humans and robots can use bi-directional communication to influence trust, adding transparency and explainability regarding their intents and capabilities to adjust expectations regarding each other.

Despite the eventual improvement on multi-task trust prediction performance, the use of task capability requirements could also be considered a drawback of our model because it calls for one more subjective input dimension in comparison with current models. Rating and describing tasks that must be executed by humans and robots in terms of specific human/robotic capability dimensions depends on the trustor agent's individual beliefs and experiences—natural, in the case of a human trustor agent, or artificial, in the case of a robotic trustor agent. Our models' trust prediction performance might have also been restricted by inconsistencies related to task char-

acterization by each participant of our experiment. We believe that better trust prediction results can be achieved with in-person longitudinal experiments involving fewer participants and more predictions.

6.6 Conclusion and Contribution

This chapter's main contribution is the proposed multi-task bi-directional trust model, which depends on both a trustee agent's proven capabilities (as observed by the trustor agent) and on the task capability requirements (as characterized by that same trustor agent). As shown in Section 6.4, our model outperforms the most relevant and recent trust models (i.e., [132] and [115]) in terms of predicting the transferred trust between distinct tasks by addressing the main limitations of those models, mostly related to a lack of task requirements descriptions. With a generalist capability dimension representing trustee agents' capabilities, our model can also represent robots' artificial trust in different trustee agents. Our model is useful for future applications where humans and robots collaborate and must sequentially take turns in executing different tasks.

CHAPTER VII

Conclusion

This dissertation investigated factors that affect drivers' trust in ADSs, methods for processing and influencing drivers' trust in ADSs, and computational models of trust. Advances in perception and artificial intelligence technology are expected to lead to seamless interaction between humans and robots in the near future. In either human-robot or driver-ADS interactions, those intelligent autonomous systems need to understand their human counter part's behaviors that reflect trust, and adapt their autonomously generated decisions taking estimates of humans' trust into consideration. The four main contributions of this dissertation are presented in detail below.

7.1 Contributions

7.1.1 Investigation and characterization of risk factors that affect drivers' trust in ADSs

In Chapter III and in [10], we explored the influence of internal and external risk on ADS trust and on how ADS trust impacts the following trusting behaviors from AV drivers: ADS monitoring and NDRT performance. We presented a 2×2 (internal vs. external risk factors) within-subjects user experiment with 37 participants that,

in summary, contributes to the literature on trust in driver-ADS interaction with the following main findings:

- Internal risk imposes limits to the expected positive impact of ADS trust and NDRT performance. In other words, trusting an unreliable ADS will not lead to better NDRT performance. However, external risk (i.e., low visibility) does not have a significant influence on ADS trust nor on how ADS trust impacts NDRT performance.

- Nonetheless, external risk was a factor that moderated the impact of ADS trust on ADS monitoring. Particularly, this means that when visibility is low, drivers are generally not able to reduce ADS monitoring, even when they report to be highly trusting the ADS.

These findings were preparatory for the development of the methods for trust estimation and trust calibration, which are the next contributions of this dissertation, described in Chapter IV and Chapter V.

7.1.2 Method for real-time trust estimation

The second contribution of this dissertation is a new real-time method for trust estimation. This method is based on a Kalman-filter approach that processed "easy-to-sense" variables, such as the drivers' focus on the NDRT (obtained with an eye-tracking device), the drivers' performance on the NDRT and the drivers' usage of the ADS self-driving functions. The method, presented in Chapter IV and in [6], brings innovations to driver-ADS trust literature, as there was a lack of practical methods for estimating drivers' ADS trust.

7.1.3 Method for trust calibration

The third contribution of this dissertation is the development of a trust calibrator that leverages the trust estimates resulting from the trust estimation method described in Chapter IV. The method for trust calibration consists of identifying trust miscalibrations (i.e., under- and overtrust) from comparisons between trust estimates and trust references. These trust references are given by the level of AV capabilities, which vary according to the driving conditions faced by the driver and the AV. Right after the identification of trust miscalibrations, different communications from the AV to the driver are triggered, with corresponding messages and styles. These different messages and styles have the goal of encouraging undertrusting drivers or warning overtrusting drivers. The combination of the trust estimator and the trust calibrator originates the trust management framework, described in detail in Chapter V and [7].

7.1.4 Bi-directional trust model

The fourth and final contribution of this dissertation is a bi-directional trust model for either predicting a humans' natural trust or determining a robot's artificial trust. This model extends the current applications of the existing trust computational models, which are mostly used for (humans' natural) trust prediction only. Moreover, those existing trust models have several limitations related to transferring trust in an agent from one task to another different task. In other words, by using the existing trust models, a robotic system is able to compute a human's trust in an agent a to execute a task γ after that human had observed that same agent a executing that same task γ. However, accurately computing a human's trust in that agent a to execute a different task γ' (which may or may not be similar to γ) is much more challenging. The bi-directional trust model proposed in this dissertation makes advances in solving this problem, as described in Chapter VI and [9].

7.2 Limitations

Although we have used techniques suitable for data with uncertainty (such as using a Kalman filter for our trust estimation), the results of our methods could certainly improve from having larger trust datasets. The relatively small size of the datasets obtained experimentally was a limiting factor of the trust estimation accuracy. Additionally, in all methods presented, the trust models had parameters that were averaged from a pool of participants rather than particularized for each participant, increasing the uncertainty of the parameters and decreasing the accuracy of trust estimates. In particular, the experiments described in Chapter III and Chapter IV were sufficient to show the effectiveness of our techniques. Still, these techniques could have a better performance if we had more data available. This limitation speaks to the difficulties involved in analyzing data that were self-reported by human subjects. Self-reported trust data based on standard questionnaires are particularly noisy and, therefore, it is extremely challenging to establish accurate ground truths for trust. Participants are not really capable of precisely estimating their own trust levels, as they are given constrained standard questionnaires that may not be able to capture all trust facets. This is not a particular limitation from our work, and it has been considered a big challenge for the human-robot interaction field [18]. A possible approach to reduce the uncertainty in both parameters and trust estimates could consist in conducting longitudinal experiments, with a lower number of participants that had more interactions with the systems in different opportunities, such as daily or weekly. This approach would allow for the establishment of particular instances of the trust model, with parameters that were optimized for each participant.

Another limitation that relates to the previous one is that the experiments of this dissertation were all implemented in simulated environments only and not in real AVs. Reproducing the risks and the nuances of driving, even in high-fidelity simulators, is challenging [22]. Because of the possible differences between having

participants in a simulated environment and in real-world situations, the conclusions and results presented in this dissertation should be considered part of an incipient body of knowledge that should be verified and validated in actual AVs. Implementing the methods proposed in this research in real vehicles would open paths for exploring new improvements and identifying possible limitations not reproducible in simulation. We alert the reader, however, that a considerable body of research has pointed out the similarities between experiments carried out in simulators and in real-world systems [39, 42, 73, 109, 120]. For that reason, we consider that our conclusions are likely to be reinforced—and not contradicted—by new experiments with real vehicles.

7.3 Future Work

The trust management framework proposed in Chapter V assumes that the AV's capability in different driving situations is quantifiable, measurable, and represented as a discrete ordinal variable (low, medium, and high). While many different methods to determine the AV's capability could be proposed, this dissertation refrains from suggesting how that should be done. For example, the AV could identify inappropriate driving conditions, such as being outside of its geofence, or being on an unsignalized road. Moreover, the AV could process the reliability of its sensors, such as the GPS processor or the inertial measurement units, which are fundamental for the dynamic driving task. Therefore, the implementation of the proposed methods outside the lab-controlled environment will depend on the prior definition of *how to determine the AV's capabilities*. A future research direction could be focused on defining a methodology for translating the limitations of self-driving functions in different situations into capability metrics and representing the corresponding trust reference levels. This research direction would require a deeper investigation of what are the main weaknesses of current ADSs, and in what situations those ADSs are not capable of sustaining the dynamic driving task [101]. Additionally, research would be

required for the definition of methods to identify these situations in real time, with the specification of possibly new sensors and procedures. The degree of difficulty involved in driving and obtained from the information provided by those sensors could be eventually reflected by a metric representing the AV capabilities.

Another research direction, more focused on the bi-directional trust model presented in Chapter VI, is to investigate methods to define and represent task requirements and agent capabilities. The bi-directional trust model was established from high-level capability representations, which were used to characterize both the agent and the task. In the presented examples, those capabilities were either arbitrarily assigned by the participants (in the case of the experiment conducted for human natural trust prediction) or randomly generated (in the case of the simulations for the definition of a robot's artificial trust). Moreover, those capabilities were considered static rather than dynamic variables. For this reason, those capability representations alone can not capture the possible evolution of the agents' knowledge or competence in executing the tasks that are assigned to them over time. As agents can typically achieve higher levels of performance in some tasks by training, or lose proficiency in some tasks after long terms without executing these tasks, those agents' capability representations must be sufficiently sophisticated so that the capability changes over time can be accurately characterized. In the future, the proposed bi-directional trust model could be extended to include not only agents' capabilities, but also their capacity, availability, situation awareness and workload. Therefore, a possible direction for additional research is to deepen the investigation on how to accurately represent both task requirements and agent capabilities, in order to improve the performance of the bi-directional trust model.

For the bi-directional trust model to be applied for solving task allocation problems between a human and a robotic agent, however, it needs to be combined with numerical representations of task rewards and costs. These rewards and costs may

depend not only on the task to be executed, but also on the agents themselves. A fundamental question in this problem is: how to define the rewards and costs associated with the execution of a given task by a given agent? Many robotic decision-making problems are focused on establishing the algorithms to optimize a certain reward or cost function, but generally ignore how those functions should be defined. Therefore, a relevant research goal is to investigate how to learn those reward/cost functions from human demonstration. This goal is very similar to that of the inverse reinforcement learning (IRL) problem, which is to extract a reward function from an observed optimal behavior [84]. The first step in this investigation could be the assessment of whether the existing IRL techniques could be applied for solving the problem of assigning rewards and costs to a set of tasks of a human-robot team.

7.4 Outlook and Impact

This dissertation deepens the knowledge on trust between humans and robotic systems, with a special emphasis on SAE level 3 AVs (and also higher SAE levels) and their human users. We present novel trust models and processing methods that are intended to advance towards the development of trust-based techniques for human-robot (driver-AV) interaction. We focus on contributing for the solutions of two high-level common problems in HRI: avoiding trust miscalibrations in driver-AV interactions and allocating tasks between a human and a robotic system that interact and work together to achieve a joint goal.

The implications of the study presented in Chapter III and the trust estimation and calibration methods presented in Chapter IV and Chapter V are all directed at solving the trust miscalibration problem. AV designers and engineers can use these methods and overall knowledge here presented to actively monitor the human users' behavior and to interact with them accordingly in order to reduce trust miscalibrations and, consequently, reduce the occurrence of accidents and improve driving

performance.

The bi-directional trust model described in Chapter VI allows for the assessment of how trustworthy a specific agent is to execute a specific task. The model is a first step for the development of a trust-based control authority allocation framework based on the following logic: as trust is fundamentally represented by a probability that the agent will succeed when executing the task, trust can be directly used in the computation of expected rewards for both the human and the robotic agent (to execute that task). The agent that *a priori* maximizes this expected reward is, therefore, the agent who should be allocated the control authority to execute the task. The outcome of that task execution should be fed back to update the parameters of the trust model representing the agents' capabilities.

APPENDICES

APPENDIX A

Risk and Trust Surveys

A.1 Post-trial Trust Survey

The following is a reproduction of the questions used in Chapter III and in [10] to measure participants' trust in the automated driving systems (ADS) after each trial, adapted from [83]. The participants were instructed to use slider bars to indicate the extent to which they believed the autonomy had each of the trust-related traits, ranging from 1 (none at all) to 7 (extremely high).

- **Competence.** To what extent did the autonomy perform its function properly? (In other words, to what extent does the driving autonomy prevent and help prevent collisions and enable safe multi-tasking?)
- **Predictability.** To what extent can the autonomy's behavior be predicted from moment to moment?
- **Reliability over time.** To what extent does the autonomy respond similarly when it encounters similar circumstances at different points in time?
- **Dependability.** To what extent can you count on the autonomy to do its job?

- **Responsibility.** To what extent did the autonomy perform the task it was designed to do? (In other words, to what extent does the driving autonomy drive safely and enable safe multi-tasking?)

A.2 Post-trial Risk Survey

The following is a reproduction of the statements used in Chapter III and in [10] to measure participants' perceived risk after each trial, adapted from [100]. The participants were instructed to place a number ranging from 1 (strongly disagree) to 7 (strongly agree) next to each statement to indicate the extent to which they agreed or disagreed.

Visibility-related statements.

- The weather made the driving situation risky.
- Due to the weather conditions, the likelihood of a collision was high.
- There was a high chance of an accident occurring because of the weather conditions.
- Due to the weather conditions, the driving situation was unpredictable.

Reliability-related statements.

- The reliability of the automated vehicle (AV) made the driving situation risky.
- Due to the reliability of the AV, the likelihood of a collision was high.
- There was a high chance of an accident occurring because of the AV's reliability.
- The reliability of the AV made the driving situation more unpredictable.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, 8(4):1–20, 11 2018.

[2] Kaarin J. Anstey, Joanne Wood, Stephen Lord, and Janine G. Walker. Cognitive, sensory and physical factors enabling driving safety in older adults. *Clinical Psychology Review*, 25(1):45–65, 1 2005.

[3] Brenna Argall and Todd Murphey. Computable trust in human instruction. In *2014 AAAI Fall Symposium Series*, 2014.

[4] Hebert Azevedo-Sa, Suresh Jayaraman, Connor Esterwood, X Jessie Yang, Lionel Robert, and Dawn Tilbury. Comparing the effects of false alarms and misses on humans' trust in (semi) autonomous vehicles. In *2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 2020.

[5] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor Esterwood, Xi Jessie Yang, Lionel Robert, and Dawn Tilbury. Real-time estimation of drivers' trust in automated driving systems. *International Journal of Social Robotics*, 2020.

[6] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T. Esterwood, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. Real-Time Estimation of Drivers' Trust in Automated Driving Systems. *International Journal of Social Robotics*, pages 1–17, 9 2020.

[7] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. Context-Adaptive Management of Drivers' Trust in Automated Vehicles. *IEEE Robotics and Automation Letters*, 5(4):6908–6915, 10 2020.

[8] Hebert Azevedo-Sa, X Jessie Yang, Lionel Robert, and Dawn Tilbury. Handling trust between drivers and automated vehicles for improved collaboration. In *2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion)*. ACM, 2021.

[9] Hebert Azevedo-Sa, X Jessie Yang, Lionel Robert, and Dawn Tilbury. A unified bi-directional model for natural and artificial trust in human-robot collaboration. *IEEE Robotics and Automation Letters*, 2021.

[10] Hebert Azevedo-Sa, Huajing Zhao, Connor Esterwood, X Jessie Yang, Dawn M Tilbury, and Lionel P Robert Jr. How internal and external risks affect the relationships between trust and driver behavior in automated driving systems. *Transportation research part C: emerging technologies*, 123:102973, 2021.

[11] Bernard Barber. *The logic and limits of trust*, volume 96. Rutgers University Press New Brunswick, NJ, 1983.

[12] Chandrayee Basu and Mukesh Singhal. Trust dynamics in human autonomous vehicle interaction: A review of trust models. In *2016 AAAI Spring Symposium Series*, 2016.

[13] Chandrayee Basu, Qian Yang, David Hungerman, Mukesh Sinahal, and Anca D Draqan. Do you want your autonomous car to drive like you? In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 417–425. IEEE, 2017.

[14] Cristiano. Castelfranchi and Rino. Falcone. *Trust theory : a socio-cognitive and computational model*. John Wiley & Sons, 2010.

[15] George Charalambous, Sarah Fletcher, and Philip Webb. The development of a scale to evaluate trust in industrial human-robot collaboration. *International Journal of Social Robotics*, 8(2):193–209, 2016.

[16] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 307–315, 2018.

[17] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Trust-Aware Decision Making for Human-Robot Collaboration. *ACM Transactions on Human-Robot Interaction*, 9(2):1–23, 2 2020.

[18] Meia Chita-Tegmark, Theresa Law, Nicholas Rabb, and Matthias Scheutz. Can you trust your trust measure? In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 92–100, 2021.

[19] Aaron Cohen. Organizational trust. In *Fairness in the Workplace*, pages 51–66. Springer, 2015.

[20] Marvin S. Cohen, Raja Parasuraman, and Jared T. Freeman. Trust in decision aids: A model and its training implications. In *in Proceedings of Command and Control Research and Technology Symposyum*, pages 1–37, Arlington, VA, 1998. Cognitive Technologies.

[21] Ewart J de Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, pages 1–20, 2019.

[22] Joost De Winter, Peter M van Leeuwen, and Riender Happee. Advantages and disadvantages of driving simulators: A discussion. In *Proceedings of measuring behavior*, volume 2012, page 8th. Citeseer, 2012.

[23] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 251–258. IEEE, 2013.

[24] Cyriel Diels and Jelte E Bos. User interface considerations to prevent self-driving carsickness. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 14–19. ACM, 2015.

[25] Cyriel Diels and Jelte E Bos. Self-driving carsickness. *Applied Ergonomics*, 53:374–382, 2016.

[26] Stephen R Dixon and Christopher D Wickens. Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human factors*, 48(3):474–486, 2006.

[27] Stephen R. Dixon, Christopher D. Wickens, and Jason S. McCarley. On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, 49(4):564–572, 8 2007.

[28] Joshua E Domeyer, Sean Seaman, Linda Angell, Joonbum Lee, Bryan Reimer, Chong Zhang, and Birsen Donmez. SHRP2 NEST database: Exploring conditions of secondary task engagement in naturalistic trip data. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 185–190. ACM, 2016.

[29] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K Pradhan, X Jessie Yang, and Lionel P Robert Jr. Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, 104:428–442, 2019.

[30] Phillip J Durst, Christopher Goodin, Chris Cummins, Burhman Gates, Burney Mckinley, Taylor George, Mitchell M Rohde, Matthew A Toschlog, and Justin Crawford. A real-time, interactive simulation environment for unmanned ground vehicles: The autonomous navigation virtual environment laboratory (anvel). In *2012 Fifth International Conference on Information and Computing Science*, pages 7–10, Shanghai, China, 2012. IEEE.

[31] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.

[32] Daniel J Fagnant and Kara Kockelman. Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77:167–181, 2015.

[33] Maziar Fooladi Mahani, Longsheng Jiang, and Yue Wang. A Bayesian Trust Inference Model for Human-Multi-Robot Teams. *International Journal of Social Robotics*, pages 1–15, 10 2020.

[34] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. Measurement of trust in human-robot collaboration. In *2007 International Symposium on Collaborative Technologies and Systems*, pages 106–114. IEEE, 2007.

[35] Christos Gkartzonikas and Konstantina Gkritza. What have we learned? A review of stated preference and choice studies on autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 98:323–337, 2019.

[36] Michael A Goodrich and Alan C Schultz. *Human-robot interaction: a survey*. Now Publishers Inc, 2008.

[37] Gregory M Gremillion, Jason S Metcalfe, Victor J Paul, and Corey Atwater. Analysis of trust in autonomy for convoy operations. In *Micro-and Nanotechnology Sensors, Systems, and Applications VIII*. Bellingham, WA: International Society for Optics and Photonics, 2016.

[38] Yaohui Guo and X. Jessie Yang. Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics*, pages 1–11, 10 2020.

[39] David Hallvig, Anna Anund, Carina Fors, Göran Kecklund, Johan G Karlsson, Mattias Wahde, and Torbjörn Åkerstedt. Sleepy driving on the real road and in the simulator—a comparison. *Accident Analysis & Prevention*, 50:44–50, 2013.

[40] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. Presenting system uncertainty in automotive uis for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*, pages 210–217. ACM, 2013.

[41] Sebastian Hergeth, Lutz Lorenz, Roman Vilimek, and Josef F Krems. Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors*, 58(3):509–519, 2016.

[42] Arsalan Heydarian, Joao P Carneiro, David Gerber, Burcin Becerik-Gerber, Timothy Hayes, and Wendy Wood. Immersive virtual environments versus physical built environments: A benchmarking study for building design and user-built environment explorations. *Automation in Construction*, 54:116–126, 2015.

[43] Kevin Hoff and Masooda Bashir. A theoretical model for trust in automated systems. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, page 115, New York, New York, USA, 2013. ACM Press.

[44] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015.

[45] Wan-Lin Hu, Kumar Akash, Tahira Reid, and Neera Jain. Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions. *IEEE Transactions on Human-Machine Systems*, pages 1–13, 2018.

[46] Y-TC Hung, Alan R Dennis, and Lionel Robert. Trust in virtual teams: Towards an integrative model of trust formation. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, pages 11–pp, Honolulu, HI, 2004. IEEE.

[47] A Hamish Jamson and Natasha Merat. Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2):79–96, 2005.

[48] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000.

[49] Siddartha Khastgir, Stewart Birrell, Gunwant Dhadyalla, and Paul Jennings. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation research part C: Emerging Technologies*, 96:290–303, 2018.

[50] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[51] Bing Cai Kok and Harold Soh. Trust in Robots: Challenges and Opportunities. *Current Robotics Reports 2020*, pages 1–13, 9 2020.

[52] Moritz Körber, Eva Baseler, and Klaus Bengler. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, 66:18–31, 2018.

[53] Tuomo Kujala. Efficiency of visual time-sharing behavior: The effects of menu structure on POI search tasks while driving. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 63–70. ACM, 2009.

[54] J. D. Lee and K. A. See. Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[55] J. D. Lee and K. A. See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, 1 2004.

[56] Jihye Lee, Daeho Lee, Yuri Park, Sangwon Lee, and Taehyun Ha. Autonomous vehicles can be shared, but a feeling of ownership is important: Examination of the influential factors for intention to use autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 107:411–422, 2019.

[57] Jiin Lee, Naeun Kim, Chaerin Imm, Beomjun Kim, Kyongsu Yi, and Jinwoo Kim. A question of trust: An ethnographic study of automated cars on real roads. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 201–208. ACM, 2016.

[58] John Lee and Neville Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.

[59] John D Lee and Neville Moray. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, 40(1):153–184, 1994.

[60] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[61] Joshua Lee, Jeffrey Fong, Bing Cai Kok, and Harold Soh. Getting to Know One Another: Calibrating Intent, Capabilities and Trust for Human-Robot Collaboration. *arXiv*, 8 2020.

[62] Monica N Lees and John D Lee. The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics*, 50(8):1264–1286, 2007.

[63] Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH Journal*, 1(1):1–14, 2014.

[64] Michael Lewis, Katia Sycara, and Phillip Walker. The role of trust in human-robot interaction. In *Foundations of trusted autonomy*, pages 135–159. Edited by Hussein A. Abbass, Jason Scholz and Darryn J. Reid, Springer, 2018.

[65] Michael Lewis, Katia Sycara, and Phillip Walker. The Role of Trust in Human-Robot Interaction. In *Studies in Systems, Decision and Control*, volume 117, pages 135–159. Springer International Publishing, 2018.

[66] Peng Liu, Run Yang, and Zhigang Xu. Public acceptance of fully automated driving: effects of social trust and risk/benefit perceptions. *Risk Analysis*, 39(2):326–341, 2019.

[67] Yidu Lu and Nadine Sarter. Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems*, 2019.

[68] Markus Maurer, J Christian Gerdes, Barbara Lenz, and Hermann Winner. *Autonomous driving*. New York, NY: Springer, 2016.

[69] Roger C. Mayer, James H. Davis, and F. David Schoorman. An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3):709, 7 1995.

[70] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[71] Natasha Merat, A Hamish Jamson, Frank CH Lai, and Oliver Carsten. Highly automated driving, secondary task performance, and driver state. *Human Factors*, 54(5):762–771, 2012.

[72] JS Metcalfe, AR Marathe, B Haynes, VJ Paul, GM Gremillion, K Drnec, C Atwater, JR Estepp, JR Lukos, EC Carter, et al. Building a framework to manage trust in automation. In *Micro-and Nanotechnology Sensors, Systems, and Applications IX*, volume 10194, page 101941U. International Society for Optics and Photonics, 2017.

[73] Lynn Meuleners and Michelle Fraser. A validation study of driving errors using a driving simulator. *Transportation research part F: traffic psychology and behaviour*, 29:14–21, 2015.

[74] Joachim Meyer. Effects of warning validity and proximity on responses to warnings. *Human Factors*, 43(4):563–572, 12 2001.

[75] Joachim Meyer. Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, 46(2):196–204, 8 2004.

[76] Abhijai Miglani, Cyriel Diels, and Jacques Terken. Compatibility between trust and non-driving related tasks in UI design for highly and fully automated driving. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '16 Adjunct, pages 75–80, 2016.

[77] Christopher A Miller. Trust in adaptive automation: The role of etiquette in tuning trust via analogic and affective methods. In *Proceedings of the 1st international conference on augmented cognition*, pages 22–27, 2005.

[78] David Bryan Miller and Wendy Ju. Joint cognition in automated driving: Combining human and machine intelligence to address novel problems. In *2015 AAAI Spring Symposium Series*, 2015.

[79] Alexander G. Mirnig, Philipp Wintersberger, Christine Sutter, and Jürgen Ziegler. A framework for analyzing and calibrating trust in automated vehicles. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Automotive UI '16 Adjunct, pages 33–38, 2016.

[80] Shane T Mueller and Brian J Piper. The psychology experiment building language (PEBL) and PEBL test battery. *Journal of neuroscience methods*, 222:250–259, 2014.

[81] Bonnie M Muir. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539, 1987.

[82] Bonnie M Muir. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11):1905–1922, 1994.

[83] Bonnie M Muir and Neville Moray. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, 1996.

[84] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[85] Brittany E. Noah, Philipp Wintersberger, Alexander G. Mirnig, and Roderick McCall. First workshop on trust in the age of automated driving. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct*, Automotive UI '17, pages 15–21, 2017.

[86] Sina Nordhoff, Miltos Kyriakidis, Bart Van Arem, and Riender Happee. A multi-level model on automated vehicle acceptance (mava): a review-based study. *Theoretical issues in ergonomics science*, 20(6):682–710, 2019.

[87] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for supervised autonomous vehicles. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '18, 2018.

[88] Ilias Panagiotopoulos and George Dimitrakopoulos. An empirical investigation on consumers' intentions towards autonomous driving. *Transportation Research Part C: Emerging Technologies*, 95:773–784, 2018.

[89] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

[90] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035, 2019.

[91] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[92] Luke Petersen, Lionel Robert, Jessie Yang, and Dawn Tilbury. Situational awareness, driver's trust in automated driving systems and secondary task performance. *SAE International Journal of Connected and Autonomous Vehicles, 2 (2)*, 2019.

[93] Luke Petersen, Dawn Tilbury, Xi Jessie Yang, and Lionel Robert. Effects of augmented situational awareness on driver trust in semi-autonomous vehicle operation. In *Ground Vehicle Systems and Engineering Technology Symposium*, 2017.

[94] Luke Petersen, Huajing Zhao, Dawn Tilbury, X Jessie Yang, Lionel Robert, et al. The influence of risk on driver's trust in semi-autonomous driving. In *In Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium (GVSETS 2018)*, pages 1–7, Novi, MI, 2018. NDIA.

[95] Vlad L Pop, Alex Shrewsbury, and Francis T Durso. Individual differences in the calibration of trust in automation. *Human factors*, 57(4):545–556, 2015.

[96] John K Rempel, John G Holmes, and Mark P Zanna. Trust in close relationships. *Journal of personality and social psychology*, 49(1):95, 1985.

[97] Nancy Rhodes and Kelly Pivik. Age and gender differences in risky driving: The roles of positive affect and risk perception. *Accident Analysis and Prevention*, 43(3):923 – 931, 2011.

[98] Lionel P. Robert. Personality in the human robot interaction literature: A review and brief critique. *Proceedings of the 24th Americas Conference on Information Systems*, pages 16–18, 2018.

[99] Lionel P. Robert, Rasha Alahmad, Connor Esterwood, Sangmi Kim, Sangseok You, and Qiaoning Zhang. A review of personality in human–robot interactions. *Foundations and Trends in Information Systems*, 4(2):107–212, 2020.

[100] Lionel P Robert, Alan R Denis, and Yu-Ting Caisy Hung. Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *Journal of Management Information Systems*, 26(2):241–279, 2009.

[101] SAE International. *SAE J3016 Standard: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. SAE International, Warrendale, PA, USA, 2018.

[102] Hamed Saeidi, John R Wagner, and Yue Wang. A mixed-initiative haptic teleoperation strategy for mobile robotic systems based on bidirectional computational trust analysis. *IEEE Transactions on Robotics*, 33(6):1500–1507, 2017.

[103] Hamed Saeidi and Yue Wang. Incorporating trust and self-confidence analysis in the guidance and control of (semi) autonomous mobile robotic systems. *IEEE Robotics and Automation Letters*, 4(2):239–246, 2018.

[104] Tracy Sanders, Alexandra Kaplan, Ryan Koch, Michael Schwartz, and Peter A Hancock. The relationship between trust and use choice in human-robot interaction. *Human Factors*, 61(4):614–626, 2019.

[105] Kristin Schaefer. *The perception and measurement of human-robot trust*. PhD thesis, University of Central Florida, Orlando, FL, 2013.

[106] Howard J Seltman. *Experimental design and analysis*. Carnegie Melon University, 2012.

[107] Kamran Shafi. A machine competence based analytical model to study trust calibration in supervised autonomous systems. In *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI)*, pages 245–252. IEEE, 2017.

[108] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017.

[109] Orit Shechtman, Sherrilene Classen, Kezia Awadzi, and William Mann. Comparison of driving errors between on-the-road and simulated driving assessment: a validation study. *Traffic injury prevention*, 10(4):379–385, 2009.

[110] Barry Sheehan, Finbarr Murphy, Cian Ryan, Martin Mullins, and Hai Yue Liu. Semi-autonomous vehicle motor insurance: A Bayesian network risk transfer approach. *Transportation Research Part C: Emerging Technologies*, 82:124–137, 2017.

[111] Shili Sheng, Erfan Pakdamanian, Kyungtae Han, Ziran Wang, John Lenneman, and Lu Feng. Trust-based route planning for automated vehicles. In *12th ACM/IEEE International Conference on Cyber-Physical Systems (with CPS-IoT Week 2021) (ICCPS '21)*. ACM, 2021.

[112] Thomas B Sheridan, Tibor Vámos, and Shin Aida. Adapting automation to man, culture and society. *Automatica*, 19(6):605–612, 1983.

[113] Missie Smith, Jillian Streeter, Gary Burnett, and Joseph L Gabbard. Visual search tasks: The effects of head-up displays on driving and task performance. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 80–87. ACM, 2015.

[114] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research*, page 0278364919866905, 2019.

[115] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research*, 39(2-3):233–249, 3 2020.

[116] Sonja Stockert, Natalie Tara Richardson, and Markus Lienkamp. Driving in an increasingly automated world–approaches to improve the driver-automation interaction. *Procedia Manufacturing*, 3:2889–2896, 2015.

[117] Nathan L Tenhundfeld, Ewart J de Visser, Kerstin S Haring, Anthony J Ries, Victor S Finomore, and Chad C Tossell. Calibrating trust in automation through familiarity with the autoparking feature of a tesla model x. *Journal of cognitive engineering and decision making*, 13(4):279–294, 2019.

[118] Jennifer E. Thropp, Tal Oron-Gilad, James L. Szalma, and Peter A. Hancock. Calibrating adaptable automation to individuals. *IEEE Transactions on Human-Machine Systems*, 48(6):691–701, 12 2018.

[119] Anne Treisman. Preattentive processing in vision. *Computer vision, graphics, and image processing*, 31(2):156–177, 1985.

[120] Geoffrey Underwood, David Crundall, and Peter Chapman. Driving simulator validation with hazard perception. *Transportation research part F: traffic psychology and behaviour*, 14(6):435–446, 2011.

[121] Frank MF Verberne, Jaap Ham, and Cees JH Midden. Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors*, 54(5):799–810, 2012.

[122] Alan R Wagner, Paul Robinette, and Ayanna Howard. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(4):1–24, 2018.

[123] Gang Wang, Bolun Wang, Tianyi Wang, Ana Nika, Haitao Zheng, and Ben Y Zhao. Defending against sybil devices in crowdsourced mapping services. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 179–191, 2016.

[124] Kirsten Weir. The dawn of social robots. *Monitor on Psychology*, 49(1):50, 2018.

[125] Darrell M West. Moving forward: Self-driving vehicles in China, Europe, Japan, Korea, and the United States. *Center for Technology Innovation at Brookings: Washington, DC, USA*, 2016.

[126] C Wickens, S Dixon, J Goh, and B Hammer. Pilot dependence on imperfect diagnostic automation in simulated uav flights: An attentional visual scanning analysis (tech rep. no. ahfd-05-02). *Urbana-Champaign, IL: Univ. of Illinois*, 21(3):3–12, 2005.

[127] Christopher D Wickens. Multiple resources and mental workload. *Human factors*, 50(3):449–455, 2008.

[128] Christopher D Wickens, Benjamin A Clegg, Alex Z Vieane, and Angelia L Sebok. Complacency and automation bias in the use of imperfect automation. *Human Factors*, 57(5):728–739, 2015.

[129] Christopher D Wickens, Stephen Rice Dixon, and Nicholas R Johnson. *UAV automation: Influence of task priorities and automation imperfection in a difficult surveillance task*. Aviation Human Factors Division, Institute of Aviation, University of Illinois at Urbana-Champaign, 2005, Chicago, IL, 2005.

[130] Christopher D Wickens, Sallie E Gordon, Yili Liu, et al. *An introduction to human factors engineering*. Longman New York, 1998.

[131] Heather Woltman, Andrea Feldstain, J Christine MacKay, and Meredith Rocchi. An introduction to hierarchical linear modeling. *Tutorials in quantitative methods for psychology*, 8(1):52–69, 2012.

[132] Anqi Xu and Gregory Dudek. OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. *ACM/IEEE International Conference on Human-Robot Interaction*, 2015-March:221–228, 2015.

[133] Anqi Xu and Gregory Dudek. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 221–228. IEEE, 2015.

[134] Rosemarie E Yagoda and Douglas J Gillan. You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *International Journal of Social Robotics*, 4(3):235–248, 2012.

[135] Xuedong Yan, Xiaomeng Li, Yang Liu, and Jia Zhao. Effects of foggy conditions on drivers' speed control behaviors at different risk levels. *Safety Science*, 68:275–287, 2014.

[136] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 408–416. IEEE, 2017.

[137] Sangseok You and Lionel P Robert. Human-robot similarity and willingness to work with a robotic co-worker. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018.

[138] Tingru Zhang, Da Tao, Xingda Qu, Xiaoyan Zhang, Rui Lin, and Wei Zhang. The roles of initial trust and perceived risk in public's acceptance of automated vehicles. *Transportation research part C: emerging technologies*, 98:207–220, 2019.

[139] Huajing Zhao, Hebert Azevedo-Sa, Connor Esterwood, X Jessie Yang, Lionel Robert, and Dawn Tilbury. ERROR TYPE, RISK, PERFORMANCE, AND TRUST: INVESTIGATING THE IMPACTS OF FALSE ALARMS AND MISSES ON TRUST AND PERFORMANCE. Technical report, Ground Vehicle Systems Engineering and Technology Symposium, 2019.