Modeling, Measuring, and Predicting Trust in Autonomous Vehicles

by

Jackie Ayoub

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial Systems Engineering)
in the University of Michigan-Dearborn
2022

Doctoral Committee:

Assistant Professor Feng Zhou, Chair
Associate Professor Shan Bao
Assistant Professor Fred Feng
Professor Wei Wang, Hunan University
Associate Professor X. Jessie Yang, University of Michigan

Jackie Ayoub

jyayoub@umich.edu

ORCID iD: 0000-0003-0274-492X

DEDICATION

I dedicate this thesis to my amazing parents, Jack and Eva Ayoub, who gave me the little they had to ensure I succeed in life and achieve my dreams. I also dedicate my work to my sisters and brother Nicol, Mireille, and Georges without their love, prayers, and encouragement I wouldn't reach this phase. I further dedicate this work to my friends who helped me out with their abilities.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Feng Zhou for imparting his knowledge, experience, and guidance throughout my research as well as life. I am thoroughly grateful for his continuous motivation and patience.

I would like to thank my committee members, Dr. Shan Bao, Dr. Jessie Yang, Dr. Fred Feng, and Dr. Wang Wei for their valuable and constructive feedback that helped me complete my research.

I would like to extend my deepest appreciation to my friends and colleagues, Lilit Avetisyan, Dania Ammar, Na Du, Doo Won Han, Jose Cordova Sanchez, and Mustapha Makki for all their help, support, and meaningful discussions we shared on life and research that have indeed made my experience more enjoyable.

I would like to give special thanks to my family and fiancé for their continuous support and understanding while undertaking my research. Your prayers were what sustained me this far.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

Although technological advances in the automotive industry are bringing autonomous vehicles (AVs) closer to road use, the public does not seem to accept AVs. One of the impeding factors to the adoption of AVs is trust. To better measure and model trust in AVs, we adopted the three layers structure, including learned trust, dispositional trust, and situational trust.

Estimating trust in AVs is challenging, especially when trust in AVs can evolve dynamically under the influence of multiple factors at the same time. For example, people's trust in AVs is shaped by whether their expectations in the capabilities of the AV system before, during, and after interaction meet the AV performance. Thus, for people's trust to match the true capabilities of the AV system and mitigate the influences of the situational factors during the interaction process, a trust calibration is needed to regulate undesirable trust levels (i.e., overtrust and undertrust). Therefore, it is essential to understand the factors affecting people's trust calibration during the human-AV interaction process.

The proposed research aims to address the previously mentioned research gaps by differentiating, measuring, modeling, and predicting the three layers of trust (i.e., learned trust, dispositional trust, and situational trust) in AVs. In Chapter 3, we studied the

influence of important factors (e.g., Benefit, Risk, Excitement, KnowledgeinAVs, EagertoAdopt) affecting people's dispositional and initial learned trust using a survey study. These factors were used as input to train an eXtreme Gradient Boosting (XGBoost) model to predict trust in AVs. To interpret the trust predictions of the XGBoost model, SHapley Additive exPlanations (SHAP) were used. Compared to traditional regression models and black-box machine learning models, our findings showed that this approach was powerful in providing a high level of explainability and predictability of trust in AVs, simultaneously. In Chapter 4, we developed a computational model to predict situational trust using physiological measurements in real time. The collected measurements were used as factors to train and test a machine learning model. The results showed that the XGBoost classifier model outperformed other machine learning models. In addition, we identified the most important physiological measures for real time prediction of trust. In Chapter 5, we studied the effect of trust precondition and system performance on the dynamic situational trust in conditional AVs. The dynamic situational trust was measured using self-reported and behavioral measures and the participants were able to adjust their self-reported situational trust levels dynamically to be consistent with the performance of the AV. However, such results were moderated by their trust preconditions measured by behavioral situational trust levels. Results showed that the participants were able to calibrate their self-reported situational trust to the real performance of the AV over time. This indicates that clearly showing the capabilities and limitations of the AV can help drivers to quickly calibrate their trust level.

The conducted studies helped in advancing our understanding of trust as a determining factor to optimize the interaction between the driver and the AV system. The

results open the path for more research on the improvement of trust prediction in real time while using more complex models and exploring more physiological data. In addition, it helps in designing a trust calibration interface by tracking the moments when the driver trust/distrust the AV in real time using physiological data and machine learning models to control drivers' trust in AVs.

CHAPTER 1

Introduction

1.1 Problem statement

The recent years have witnessed a rapid emergence of AVs research and automotive companies are increasing their interest in investing more in this technology. Research has shown that people had high expectations about AV technology, but they were still concerned about adopting it (Schoettle & Sivak, 2016). The main barrier to the adoption of AVs is the lack of public trust which is affected by the increasing reports of AV failures. Therefore, to increase AVs acceptance, people need to have an appropriate level of trust to interact appropriately with the system. For instance, one of the leading causes of recent AV accidents (e.g., Tesla's fatal crash in Florida and Uber AV crash in Arizona) (Rice, 2019; Kohli & Chadha, 2020) was the drivers' overtrust in their AVs. These accidents have caused a negative first impression on people's opinion about AV safety and capabilities.

Evaluating people's trust in AVs faces many issues. First, the majority of the public does not have any interaction experience with AVs. Second, during the interaction with an AV, people's trust can evolve dynamically under the influence of multiple factors at the same time. Third, after

experiencing the AV system, undesirable levels of trust (i.e., overtrust and undertrust) can potentially diminish the benefits of AVs leading to accidents.

A substantial amount of research has been conducted to understand the factors affecting people's trust (Numan, 1998; Kim et al., 2008; Pavlou, 2003; Choi & Ji, 2015; Bearth & Siegrist, 2016; Parasuraman & Riley, 1997; Raue et al. 2019). Researchers have identified three areas of trust variability in automation including 1) human-related area (i.e., culture, age, gender, experience, and knowledge about AVs), 2) automation-related area (i.e., reliability, uncertainty, workload, and user interface), and 3) environmental-related area (i.e., risk, brand) (Hoff & Bashir, 2014; Ayoub et al., 2019). These three areas reflect the three layers of trust: dispositional, situational, and learned trust (Hoff & Bashir, 2014). Although these three layers are dependent on each other, it is necessary to differentiate and isolate each layer to build an understanding of its effect on people's trust in AVs.

With the increasing complexity of the AV system, the driver can no longer predict when the system is not operating properly. Overtrusting the AV system can lead the driver to depend on the AV system even when it falls short of expectations. And undertrust the AV system can lead to diminishing the benefits of the system. Thus, it is necessary to model and predict drivers' trust changes over time by considering the influence of the factors affecting the three layers of trust. But before predicting people's trust in AVs, it is important to quantify their levels of trust before and after interacting with the AV system using objective (i.e., physiological measures) and subjective measures (self-reported and behavioral measurements).

1.2 Research objectives

We plan in this research to build an understanding of the three layers of trust in the AVs domain as shown in Figure 1.1. In our first study (i.e., see chapter 3), to understand the baselines to form people's trust in AVs, we measured dispositional and initial learned trust. In addition, we used machine learning models to predict drivers' dispositional and initial learned trust using survey data by identifying the factors affecting trust. In our second study (i.e., see chapter 4), we measured situational trust using physiological measurements in real time.

Also, the collected measurements were used to train a machine learning model on predicting people's trust in real time. In our third study (i.e., see chapter 5), we measured dynamic situational trust using self-reported and behavioral measures. Specifically, we investigated the effect of trust precondition (i.e., undertrust and overtrust) and system performance on the dynamic situational trust in conditional AVs, where the results provided some implications for designing an in-vehicle trust calibration system.

The objectives of the proposed research are the following:

1) Develop a computational model to predict drivers' dispositional and initial learned trust using machine learning and survey data.
2) Develop a computational model to predict situational trust using physiological measurements in real time in conditional AVs.
3) Investigate the effect of system performance and people's trust preconditions on the dynamic situational trust during takeover to provide implications for designing an alerting system to calibrate people's trust in conditional AVs.

This dissertation is presented in six chapters as described below:

Chapter one is an introductory chapter that includes the problem statement and research objectives of this work.

Chapter two provides a literature review discussing the topic of trust in AVs, factors influencing trust in AVs, and existing methods of estimating and calibrating trust in AVs.

In **Chapter three,** we measured dispositional and initial learned trust to understand the baselines to form people's trust in AVs.

In **Chapter four**, we predicted people's situational trust in real time using physiological data and a machine learning model.

In **Chapter five**, we investigated how system performance and people's trust preconditions affect their dynamic situational trust.

And finally, **Chapter Six** summarized the findings from previous chapters and provided suggestions for future research.

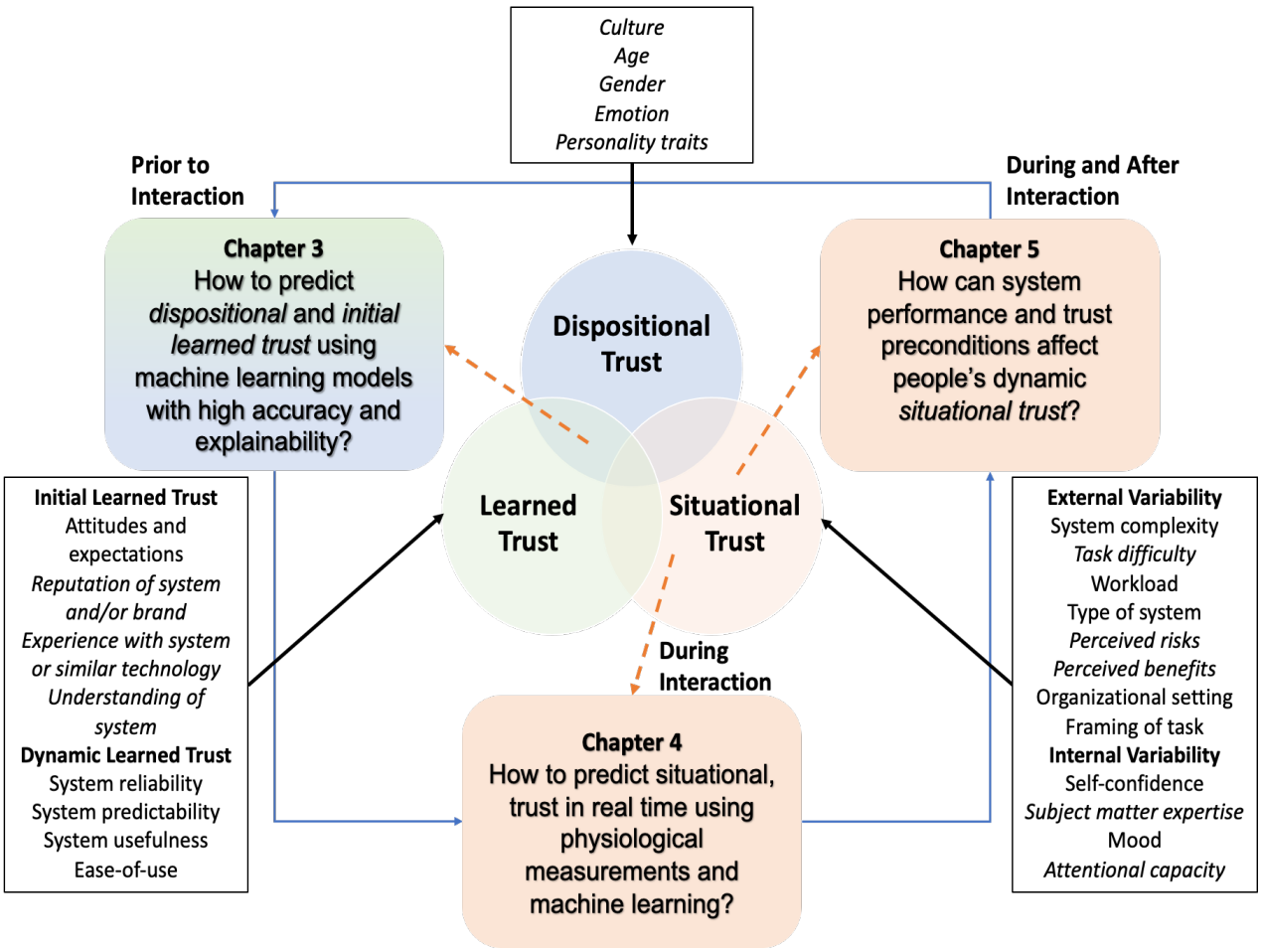


Figure 1.1: Research scope.

CHAPTER 2

Literature Review

2.1 Research background

2.1.1 Trust in autonomous vehicles

Not surprisingly, the concept of trust in automation has gained a lot of attention due to its importance in the domain of AVs. Mayer et al. (1995) have identified three foundations of trust including integrity, benevolence, and ability. Integrity and ability are both affected by the system performance whereas benevolence is affected by how much the system meets the expectations of the truster. Thus, this confirms that trust formation is a dynamic process. Trust changes drastically at both the unconscious and conscious levels as people are exposed to new information (Hoff & Bashir, 2014; Ayoub & Zhou, 2020). According to Lee & See (2004), trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". Based on this definition, the three critical elements influencing trust are the effect of information display, the closed-loop dynamics of reliance and trust, and the importance of the context on trust (Lee & See, 2004).

Hoff & Bashir (2014) defined three layers that structured the human-automation trust: learned trust, dispositional trust, and situational trust. Learned trust is influenced by people's

experience with automation which can be affected by the automation performance as well as people's knowledge about automation. It can be divided into two categories: initial learned and dynamic learned trust. Initial learned trust is related to our knowledge before interacting with the system. Whereas dynamic learned trust is built after interacting with the system based on its performance. Dispositional trust reflects peoples' rational tendency to trust automation which is affected mainly by age, culture, gender, and personality traits. Situational trust is dependent on the situation, and it can be affected by internal and external variabilities. External variabilities are affected by the system complexity, task difficulty, people's workload, perceived risks, and benefits. However, the internal variabilities are affected by people's mood, self-confidence, and expertise. Therefore, measuring people's dynamic trust while using an automated system is complex and unclear since it is affected by multiple interacting variables.

Trust in AVs research has covered a variety of different study contexts. Lee et al. (2016) showed that people's trust in AVs can be improved by providing a continuous evaluation of its performance. In addition, introducing knowledge of the AV system limitations and capabilities was shown to increase people's trust in the system (Shariff et al., 2017; Khastgir et al., 2018; Dikmen & Burns, 2017). Ayoub & Zhou (2020) showed that the trust and risk level of AV decisions were inversely proportional. In addition, they showed that trust in visual displays was higher than in tabular displays form of the system decision. Dzindolet et al. (2003) showed that a transparent system that provides useful feedback and information about why a failure might occur can increase people's trust. Furthermore, the timing of the system failure has a negative effect on trust. For instance, Manzey et al. (2012) and Sanchez (2006) showed that system failures occurring at the beginning of interaction have a greater negative impact on trust. Johnson (2012) showed that distinct types of automation failures can have different effects on people's trust behavior. In

particular, false alarms reduce people's compliance whereas misses reduce people's reliance (Davenport & Bustamante, 2010; Sanchez, 2006; Rice, 2009).

2.1.2 Factors influencing trust in autonomous vehicles

To increase the public usage of AVs, it is essential to understand the factors affecting people's trust perception. Many researchers have consistently reported the effects of risks, benefits, knowledge, and feelings on trust (Walker et al., 2016; Raue et al., 2019; Rudin-Brown & Parker, 2004; Parasuraman & Miller, 2004).

Perception of Risks: Risk is considered to be an intrinsic aspect affecting trust, i.e., when the perceived risk of a situation is high, a higher level of trust is needed to rely on AV's decisions (Numan, 1998; Kim et al., 2008; Pavlou, 2003). Therefore, it is essential to consider factors associated with risks in AVs when evaluating trust (Rajaonah et al., 2008). Zmud et al. (2016) reported that safety risks due to system failures were the major concerns of using AVs. Moreover, Menon et al. (2016) showed that one third of US drivers were worried about the risks of misusing their private AV data. Li et al. (2019) demonstrated that the perceived risks and trust in an AV were affected by introductory information related to system reliability. Therefore, it is important to include risk perception and an appropriate level of information regarding AVs to evaluate trust in the early stages of driver-vehicle interactions.

Perception of Benefits: Many researchers have found that the perception of benefits is related to improving trust in AVs, which subsequently leads to user acceptance and adoption (Choi & Ji, 2015; Bearth & Siegrist, 2016). One of the major benefits associated with AVs is to reduce vehicle crashes and to save lives. Vehicle crashes lead to injury of 2.2 million Americans each year (NHTSA, 2010) and the cost associated with these crashes is around $300 billion (Bearth &

Siegrist, 2016). Therefore, the safety enhancement behind AVs should be focused on creating crash-less vehicles (Johnson, 2012; Fagnant & Kockelman, 2015; Paden et al., 2016). As a matter of fact, human factors were reported to be the cause of 90% of crashes and the death of over 30 thousand Americans per year (Elrod, 2014). AVs are accurate and quicker to react in case of an emergency since they can optimize the decision before taking any actions. Aside from improving safety, AVs can bring other social benefits, including reducing congestions, fuel consumption, and CO2 emission (Fagnant & Kockelman, 2015), and so on.

Knowledge about AVs: Another important factor influencing trust is the knowledge of the public regarding the capabilities and limitations of AVs. A lack of knowledge in automation leads to mistrust or overtrust of the true capabilities of the system (Parasuraman & Riley, 1997). Doney et al. (1998) presented a direct effect of knowledge on trust, where knowledge reduced uncertainty which in return increased trust. Khastgir et al. (2018) demonstrated that providing introductory knowledge about AVs to the participants increased their level of trust in the system. To calibrate trust, the authors suggested the concept of information safety to ensure safe interaction with AVs. Holmes (1991) argued that trust developed with the accumulation of knowledge from increasingly more experience from the past. Therefore, experience plays an important role in shaping our trust assessment. For instance, Ruijten et al. (2018) demonstrated that mimicking human behavior using intelligent user interfaces improved drivers' trust in AVs. Edmonds (2019) showed that participants who had advanced driver-assistance systems (ADAS) in their vehicles were 68% more likely to trust these features than the drivers who did not have them.

Effect of Feelings: Trust is composed of two components: a cognitive component and an affective component (Lewicki & Brinsfield, 2011; Cho et al., 2015). The cognitive component is based on judgments, beliefs, competence, stability, and expectations while the affective component

is based on positive and negative emotions that shape our trust (Lewis & Weigert, 1985). For example, positive emotions were found to improve takeover performance in AVs, which further led to trust in AVs (Du et al., 2020) while negative emotions, such as concerns and worries, made parents trust automated school buses less (Ayoub et al., 2020). Furthermore, Peters et al. (2006) explained that affect influenced our stored knowledge, which further guided our acceptance and trust. Hence, emotion can be used to evaluate trust. According to Hancock et al. (2019), the majority of drivers had no chance to experience AVs yet. Thus, this inexperience makes it harder to evaluate their trust in the system. Raue et al. (2019) suggested that feelings related to people's experience in driving could shape their perception of risks, benefits, and trust in AVs. Specifically, Baumeister et al. (2001) showed that negative emotions were more significant in shaping judgment than positive ones.

2.1.3 Existing methods of estimating trust in autonomous vehicles

Many researchers used questionnaires (Körber, 2018) and behavioral methods (Miller et al., 2016; Jessup et al., 2019) to evaluate trust in AVs. For instance, Körber (2018) built a multidimensional model to measure trust in automation using a survey study. The model was composed of 19 parameters, including reliability, understandability, propensity to trust, familiarity, and intentions. Miller et al. (2016) developed a survey to study trust in automation by focusing on 5 components of trust including competence, predictability, dependability, consistency, and confidence. Furthermore, Lee & See (2004) summarized the factors affecting trust in automation into a three-dimensional model, including performance, process, and purpose. Jian et al. (2000) built a scale system to measure trust using an experimental study that explored the similarities and differences between trust and distrust in automation. Raue et al. (2019) used linear regression to model interests in using AVs and logistic regression to model parents' attitudes

toward children riding in AVs alone. Both models identified significant factors (e.g., risk perception, benefit perception, negative emotions in manual driving) influencing the dependent variables, but no prediction results were reported. Commonly, trust models are modeled using a linear combination of the input factors, which identify significant factors that influence trust in AVs and other automation systems. However, they did not report prediction results. Machine learning techniques were proposed in modeling trust in AVs. For example, Liu et al. (2011) investigated the usage of two machine learning models: linear discriminant analysis for feature importance and decision trees for classification for large-scale systems (e.g., product recommendation systems, Internet auction sites) with false rates between 10% and 19%. Guo et al. (2020) developed a personalized trust prediction model based on the Beta distribution and learned its parameters using Bayesian inference. López & Maag (2015) designed a generic trust model capable of processing various trust features with an SVM technique. On their simulated trust dataset, they obtained 96.61% accuracy.

2.1.4 Predicting drivers' trust state through psychophysiological measurements

To understand how trust can impact the use of automation, it is necessary to use reliable trust metrics. The majority of the existing metrics are based on questionnaires such as the trust scale suggested by Jian et al. (2000) which was used in multiple applications to measure general trust (Dzindolet et al., 2003; Gold et al., 2015; Hoff & Bashir, 2014). However, this scale did not support the temporal and context related nature of trust. Recently, Holthausen et al. (2020) suggested a short scale based on the trust model suggested by Hoff & Bashir (2014) to assess situational trust using six items including trust, performance, NDRT, risk, judgment, and reaction. However, self-reports cannot capture real-time changes in trust specifically during real-world driving (Hergeth et al., 2016, Walker et al., 2019). Many researchers have argued that

psychophysiological measurements such as galvanic skin response (GSR), gaze behavior, heart rate and electroencephalography (EEG) are the solution to objectively measure trust in real time (Hergeth et al., 2016, Walker et al., 2019, Akash et al., 2018).

GSR is the measure of people's sweat-gland activity. When there is sweat our skin starts changing its conductance which has been found to vary linearly with respect to emotional arousal (Baig & Kavakli, 2019). GSR has been used in measuring anxiety, cognitive workload, and stress (Hergeth et al., 2016; Akash et al., 2018; Jacobs et al., 1994; Hu et al., 2016). In addition, researchers found a correlation between GSR and human trust. For instance, Khawaji et al. (2015) showed that trust and cognitive workload significantly affected the GSR peak values. Kumar Akash et al. (2018) showed that the GSR peak values were significantly affected by interpersonal trust and the difficulty index of decision-making. Walker et al. (2019) showed that the higher the trust in the system, the more attention they paid to secondary tasks and the less they checked the road, and the lower the GSR. And combining GSR with gaze behavior led to a better prediction of trust. Wang et al. (2018) showed that GSR and gaze behavior were negatively associated with self-reported trust. Gaze behavior is one of the frequently used indicators of trust. For instance, drivers with a lower frequency of monitoring the road had a significantly higher trust in the AV (Hergeth et al., 2016).

EEG corresponds to the electrical activity of the brain (Baig & Kavakli, 2019). EEG has the potential to identify different arousal levels as well as differentiating between positive and negative emotional valence. Akash et al. (2018) developed an empirical trust model of object detection in AVs using a quadratic discriminant classifier and physiological measurements such as EEG and GSR. Dong et al. (2015) proposed a method to measure human-machine trust using EEG during a theory of mind game. Wang et al. (2018) identified four brain regions responsible

for human trust in automation using EEG which allows for real time assessment of human trust in automation.

2.1.5 Existing methods in calibrating trust in autonomous vehicles

To benefit from the advantages of using AVs, it is necessary to have a calibrated level of trust in the AV system. Trust calibration is defined as “the correspondence between the person’s trust in the automation and the automation’s capabilities" (Lee & See, 2004). Thus, AV trust calibration means adjusting people’s trust level based on the actual reliability of the AV system (Lee & See, 2004; Muir, 1994). Since the AV system reliability is affected by many environmental factors, people usually end up overtrusting or undertrusting the system which results in serious safety issues (Parasuraman & Riley, 1997; Robinette et al., 2016). Overestimating the reliability of the AV system leads to misusing the AV whereas underestimating the AV system leads to disusing the AV.

Okamura & Yamada (2020) proposed a trust calibration method based on people’s behavior. Once an overtrust or undertrust behavior was detected, trust calibration cues were presented to the drivers to warn them to recalibrate their trust. Whereas, Merritt et al. (2014) proposed a continuous trust calibration by displaying information about the performance of the AV. Also, Seppelt (2009) suggested a continuous trust calibration by providing information about the AV performance as well as the AV limitations. Parasuraman & Miller (2004) showed that the norms of the human-human etiquette affected the trust calibration. Automation etiquette was referred to as the explanation style that was either patient and non-interruptive or impatient and interruptive.

To improve trust calibration, Wiegmann et al. (2001) proposed the usage of interface designs (i.e., automation aids). Rovira et al. (2007) showed that people’s first-time exposure to

poor automation performance can lead to appropriate trust calibration later on. In addition, Gempler (1997) suggested that to optimize the use of automation, people's trust calibration should be along the line of Figure 2, where the perceived reliability is equal to the actual reliability. In the overtrust region, a misuse of the automation will happen since people's trust is higher than the automation reliability. However, in the undertrust region, a disuse of the automation will happen since people's trust is lower than the automation reliability.

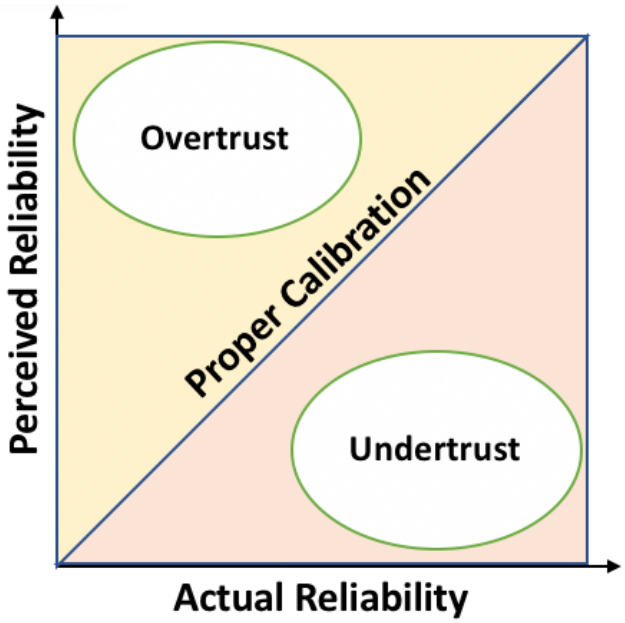


Figure 2.1: Reliability calibration.

2.2 Research objectives

The proposed research will consist of three phases. Phase 1 involves the development of a computational model to predict drivers' dispositional and initial learned trust using machine learning and survey data; phase 2 will develop a computational model to predict drivers' situational trust using physiological measurements in real time; finally, phase 3 investigates the dynamic situational trust using self-reported and behavioral measures. In addition, the results of this study

provided implications for designing an alerting system to recalibrate people's trust in AVs whenever an undertrust or overtrust situation is detected.

In my first year of the Ph.D. program, we conducted a literature review to gain a firm understanding of the research evolution from manual to automated driving. An overview of the ten-year development of the papers presented at the International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutoUI) from 2009 to 2018 (Ayoub et al., 2019) was conducted. We identified various topics and described their development related to manual driving (78.3%) and automated driving (21.7%) as shown in Table 2.1. The main identified topics for automated driving consisted of takeover (5.1%), trust and acceptance (4.3%), interacting with road users (2.5%), user interfaces (2.5%), and methodology (1.4%). The topic of trust and acceptance attracted our attention specifically that it is one of the key factors affecting people's usage of AVs.

A major challenge facing the AVs industry is whether the driver is willing to take an AV in the first place, i.e., trust in automated driving. It is well-known that an appropriate level of trust is essential for the driver to interact with the automated vehicle successfully (Ekman et al., 2018). The recent and ongoing trend is to identify various factors influencing trust in automation and how they can be manipulated to obtain an appropriate level of trust. This is also recognized by studies outside the AutoUI proceedings. For example, Lee & See (2004) identified individual, organizational, cultural, and contextual factors that influence trust in the process. Hoff & Bashir (2014) integrated both personal and system-related factors of trust by examining extensive literature. These identified factors offer guidelines to create an appropriate level of trust in user interface design. However, neither of them addresses the practical issue of estimating driver trust in real-time to manage and build an appropriate level of trust dynamically. Previous studies (Jonker

& Treur, 1999; Akash et al., 2017) have attempted to predict trust based on experience or self-reported data using dynamic models. Nevertheless, it is intrusive and not practical to request the driver to report his/her trust level during the human-machine interaction process. Also, computational models have been proposed to estimate trust in other domains. For example, Hoogendoorn et al. (2008) proposed a trust computational model based on the personal attributes of users. In this respect, we need to figure out how to make use of computational models to predict driver trust in real time to calibrate driver trust in automated driving.

Table 2.1: The major research topics identified based on the overview of the ten-year development of the papers presented at the International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutoUI) from 2009 to 2018.

Manual Driving (78.3%)	**Automated Driving (21.7%)**
User Interface (i.e., visual, auditory, haptic, gestural, olfactory, and multimodal) (36.2 %)	Takeover (5.1%)
Driver State (i.e., distraction, cognitive workload, and emotion) (14.1%)	Trust and acceptance (4.3%)
Methodology (i.e., design methods, measurement techniques, and test protocols) (9.4%)	Interacting with road users (2.5%)
Augmented Reality and Head-up displays (5.4%)	User Interfaces (2.5%)
Navigation (4.0%)	Methodology (1.4%)
Others (i.e., eco-driving, infotainment, user acceptance, and cultural differences) (9.1%)	Others (i.e., collaborative driving, non-driving related tasks, remote driving, driving styles, legal issues, culture differences) (5.8%)

CHAPTER 3

Modeling Dispositional and Initial Learned Trust in AVs with Predictability and Explainability

3.1 Introduction

Although SAE Level 3 (e.g., Audi A8 Traffic Jam Pilot) automation is responsible for the driving task most of the time and allows drivers to do non-driving related tasks (SAE, 2018), the public is reluctant to adopt the technology. A survey study showed that only 37% of their participants would probably buy an AV (Power, 2012). Menon (2015) showed that 61.5% of Americans were not willing to use AVs. As previously mentioned, the key barrier to the adoption of AVs is trust (Shariff et al., 2017, Bansal et al., 2016). Researchers identified many factors affecting peoples' trust in AVs. For instance, Ayoub et al. (2019) summarized the factors affecting trust into three categories, including 1) human-related factors (i.e., culture, age, gender, experience, workload, and knowledge about AVs), 2) automation-related factors (i.e., reliability, uncertainty, and user interface), and 3) environmental-related factors (i.e., risk, reputation of original equipment manufacturers). However, estimating trust in AVs is challenging, especially when the majority of the public does not have much interaction experience with AVs. Raue et al. (2019) suggested that peoples' experience in manual driving should potentially shape their trust

assessment in AVs. Along the same line, Abe et al. (2017) made use of manual driving characteristics (e.g., speeds and time headway) to investigate driver trust in automated driving in terms of overtaking and passing patterns. Machine learning techniques were proposed for modeling trust in AVs. Such models were able to predict people's trust in AVs to a large extent by aggregating numerous factors. However, the relative importance in predicting trust in AVs tends to be not obvious in such black-box models. Factors such as the perception of risks and benefits, feelings, and knowledge of AVs can provide an indicator of drivers' trust in AVs. These factors were used as input to train an eXtreme Gradient Boosting (XGBoost) model to predict trust in AVs. With the help of SHapley Additive exPlanations (SHAP), we were able to interpret the trust predictions of XGBoost to further improve the explainability of the XGBoost model. Compared to traditional regression models and black-box machine learning models, our findings showed that this approach was powerful in providing a high level of explainability and predictability of trust in AVs, simultaneously.

3.2 Method

The system architecture of this work is illustrated in Figure 3.1 with the following steps:

(1) Data Collection: We collected a dataset using an online survey on Amazon Mechanical Turks (AMTs). The survey was developed in Qualtrics and was integrated in AMT to collect participants' responses.

(2) Data Cleaning: We reviewed the participants' responses and removed invalid data.

(3) XGBoost Model Construction: We used a 10-fold cross validation process to optimize the parameters of XGBoost to train the model.

(4) XGBoost Model Evaluation: To evaluate the performance of the XGBoost model, we compared it with a list of machine learning models using various performance metrics, including accuracy, receiver operator characteristics area under the curve (ROC_AUC), precision, recall, and F1 measure.

(5) SHAP Explanation: In order to improve the explainability of the XGBoost model, SHAP was used to explain the model predictions both globally and locally.

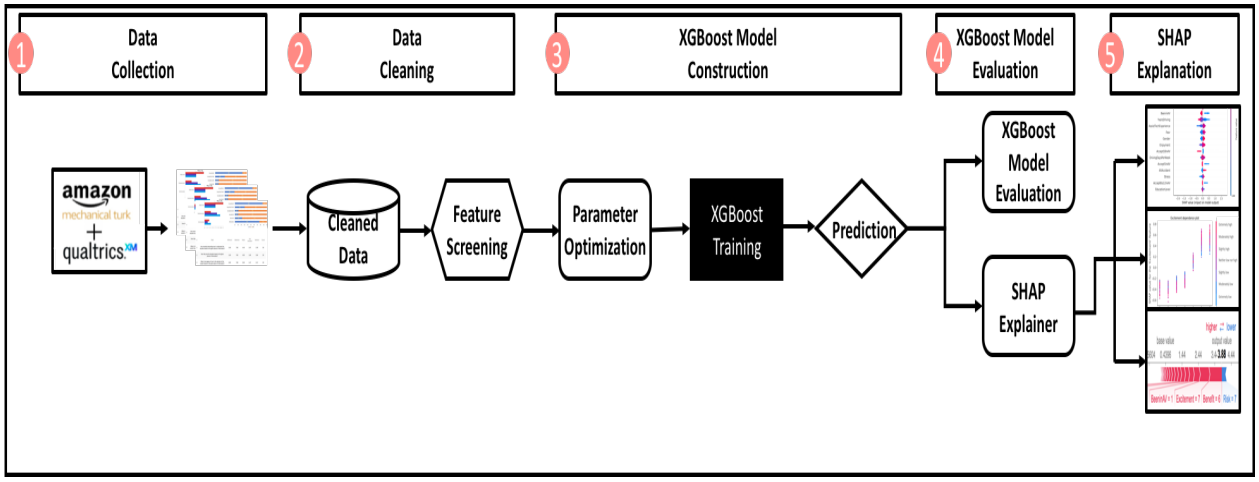


Figure 3.1: Flow chart of the proposed system architecture to predict Trust.

3.2.1 Participants and apparatus

A total number of 1175 participants located in the United States took part in the online survey using AMTs (Seattle, WA, www.mturk.com/). AMT is a web-based survey company, operated by Amazon Web Services, which has recently become popular in fast data collection (Paolacci et al. 2010). The questionnaire was developed in Qualtrics (Provo, UT, www.qualtrics.com), a web-based software to create surveys. Participants who gave nonsensical answers (i.e., unreasonable driving experience compared to their age, using letters instead of numbers to represent the number of driving years, using the same pattern to answer all the questions, and completing the survey too quickly) were excluded from the study. After the screening, we had a total number of 1054 participants (47.5% females, 52.2% males, and 0.3% others). The age distribution and the education distribution of the participants are shown in Table 3.1. Participants were compensated with $0.2 upon completion of the survey. The study was approved by the Institutional Review Board at the University of Michigan.

Table 3.1: Age and education distribution of the participants in the study.

Age Distribution	<18	18-24	25-34	35-44	45-54	55-64	>=65
	0.1%	8.3%	37.7%	22.7%	14.4%	10.9%	5.9%
Education Distribution	Professional degree	Doctoral degree	Master's degree	Bachelor's degree	Some college	Associate degree	High school degree or less
	1.2%	0.9%	18.3%	43.3%	16.9%	11.5%	7.9%

3.2.2 Survey design

We investigated various factors associated with AVs, including knowledge, experience, feelings, risk and benefit perceptions, and behavioral assessment to predict trust using a survey study. The survey questions were adapted from (Raue et al., 2019; Jian et al., 2000) and are shown in Table 3.2. Participants' knowledge about AVs was measured using their eagerness level to adopt new technology, knowledge level about AVs, and knowledge about AV crashes. Experience questions were related to the experience of using ADAS and the experience of trying AVs. As for Benefit and Risk related questions, participants had to assess how beneficial and risky the AVs were. In regard to the behavioral assessment related questions, participants were asked if they would let a child under 5 years old, between 6 and 12 years old, between 13 and 17 years old, and above 18 years old use an AV alone. Since the majority of the public had no experience in AVs yet, we asked them to rate their feelings (i.e., Control, Excitement, Enjoyment, Stress, Fear, and Nervousness) based on their experience in manual driving. Among all the items in the survey, those related to knowledge and experience directly measured participants' initial learned trust while others measured their dispositional trust. We provided abbreviated names for the survey questions to use them throughout the paper as shown in Table 3.2.

3.2.3 XGBoost model construction

XGBoost classifier was selected for predicting perceived trust in AVs (Chen and Guestrin, 2016). The boosting algorithm combines multiple decision trees into a strong ensemble model and reduces the bias by reducing the residual error at each iteration where each decision tree learns from the previous one. This process is done by adjusting the weights of decision trees while iterating the model sequentially. More accurate decision trees are given more weights. XGBoost

implements the same boosting technique with an additional regularization term. During the optimization process, an optimal output value for each tree is obtained by iteratively splitting each tree to minimize its objective function.

To build a tree, the process follows the exact greedy algorithm where it starts with all the training examples, and then it calculates the split loss reduction or gain for the root of the tree. Once the gain for all the splitted trees is calculated, the tree with the maximum gain is considered as the optimal split. The gain value should be positive in order for the selected tree to continue growing. After building the trees, pruning is performed to remove the sections with low effect on the classification. Then, an output value is calculated for each leaf which will be used to make predictions. Using these predictions, the same described process is used to build a second tree. The XGBoost algorithm combines both software and hardware optimization abilities, which result in a great performance with less computational resources by performing parallel computing. In this research, we removed the highly correlated predictor variables before starting the training process in XGBoost using the Pearson correlation coefficient. The correlation coefficient was high between age and number of driving years (0.88) and between fear and nervousness (0.87). Therefore, age and nervousness were removed. We defined the response variable as a binary one, (i.e., trust = 1 (extremely high, moderately high, and slightly high), sample size = 624, and distrust = 0 (extremely low, moderately low, and slightly low), sample size = 430) by converting its 7-point Likert scale.

In the next step, we trained the XGBoost classifier with 10-fold cross-validation to optimize the accuracy of the prediction using a randomized search for hyperparameters. The learning objective used in this study was reg: logistic regression. After we constructed the model, we compared XGBoost with other machine learning models using various performance metrics,

including accuracy, area under the receiver operating characteristics curve (ROC_AUC), precision, recall, and F1 measure. Accuracy is the fraction of corrected prediction samples divided by the total samples. ROC plots the true positive rate against the false positive rate at various threshold settings, and ROC_AUC measures the performance of a classifier in distinguishing between the two classes. Precision is defined as true positive/(true positive + false positive), recall as true positive/(true positive + false negative), and F1 measure as the harmonic mean of precision and recall, i.e., 2*precision*recall/(precision+recall) (Zhou et al., 2017).

Table 3.2: Survey questions, categories, and scale.

Categories	Survey Questions	Abbreviation	Scale
General	1) What is your gender?	Gender	
	2) What is your age?	Age	
	3) What is the highest level of school you have completed or the highest degree you have received?	EducationLevel	
	4) Do you have a valid driving license?	DrivingLicense	
	5) For how many years have you been a driver?	YearsDriving	
	6) On average, how many days a week do you drive?	DrivingDaysPerWeek	
Knowledge	7) What is your eagerness level to adopt new technologies?	EagertoAdopt	From 1 (extremely low) to 7 (extremely high)
	8) What is your knowledge level in regard to autonomous vehicles?	KnowledgeinAVs	From 1 (extremely low) to 7 (extremely high)
	9) Have you heard any stories about autonomous vehicles being involved in accidents?	AVAccident	Yes / No
Experience	10) Please indicate how much experience you have with vehicle driving assistance technology (for example: cruise control, adaptive cruise control, parking assist, lane keeping assist, blind spot detection, or others)	AssistTechExperience	From 1 (extremely low) to 7 (extremely high)
	11) Have you ever been in an autonomous vehicle?	BeeninAV	Yes / No
Benefit and risk perception	12) What is the risk level of using an autonomous vehicle?	Risk	From 1 (extremely low) to 7 (extremely high)
	13) How beneficial it is to use an autonomous vehicle?	Benefit	From 1 (extremely low) to 7 (extremely high)
Behavioral assessment	14) Would you let a child who is under 5 years old use an autonomous system alone?	Assess5inAV	Yes / No
	15) Would you let a child who is between 6 and 12 years old use an autonomous system alone?	Assess6to12inAV	
	16) Would you let a child who is between 13 and 17 years old use an autonomous system alone?	Assess13to17inAV	
	17) Would you let an adult who is above 18 years old use an autonomous system alone?	Assess18inAV	
Feelings	18) How much do you feel in control (for example: attentive, alert) when you are driving?	Control	From 1 (extremely low) to 7 (extremely high)
	19) How much do you feel excited when you are driving?	Excitement	
	20) How much do you enjoy driving?	Enjoyment	
	21) How much do you feel stressed when you are driving?	Stress	
	22) How much do you feel scared when you are driving?	Fear	
	23) How much do you feel nervous when you are driving?	Nervousness	
Trust	24) In general, how much would you trust an autonomous vehicle	Trust	From 1 (extremely low) to 7 (extremely high)

3.2.4 Explaining XGBoost model using SHAP

Shapley value is a method from coalitional game theory (Shapley, 1953), in which each player is assigned with payouts depending on their contribution to the total payout when all of them cooperate in a coalition. In our study, in the case of XGBoost model, each feature (i.e., predictor variables in XGBoost) has its fair contribution to the final prediction of trust perception on AVs. Predicting if one participant trusts or distrusts AVs can be considered as a game, and the gain in this game is the actual prediction for this participant minus the average prediction for all the participants' data. For example, if we use three feature-value sets, i.e., Benefit = 7, BeeninAV = 1, and KnowledgeinAVs = 7 to predict trust in AVs, the predicted Trust is 7 and if we use Benefit = 7 and KnowledgeinAVs = 7 to predict trust in AVs, the predicted Trust is 5. Assuming we want to calculate the Sharply value of the feature-value set, BeeninAV = 1, the contribution from the above example is 7 - 5 = 2 in trust prediction. However, this is only one coalition, we need to repeat the same process for all the possible coalitions and obtain the average of all the marginal contributions. Mathematically, the Shapley value of a feature-value set is calculated as follows (Shapley, 1953):

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} (v(S \cup \{i\}) - v(S)), \quad (3.1)$$

where n is the total number of features, S is a subset of any coalition of the features N, where the summation extends over all subsets S of N that do not contain feature i, and $v(S)$ is the contribution of coalition S in predicting trust in our study. The difference between the trust prediction and the average trust prediction is fairly distributed among all the feature-value sets in the data. Therefore, it has a solid theory in explaining machine learning models.

One limitation is that when the number of features increases (so is the exponential number of coalitions), the computation needed will be exponentially expensive. According to game theoretically optimal Shapley values, Lundberg and Lee (2017) and Lundberg et al. (2020) proposed an efficient method to calculate SHAP values, especially for tree-based models, such as XGBoost. Therefore, we can use SHAP to explain XGBoost both globally and locally. Globally, we can study how SHAP values rank the features based on their importance, how SHAP values change with regard to different feature-value sets, and how one feature interacts with another. Locally, we can explain individual predictions. Among them, the interaction effect is defined as the additional combined feature effect minus individual main feature effects:

$$\varphi_{i,j}(v) = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(n-|S|-2)!}{n!} (v(S \cup \{i,j\}) - v(S \cup \{i\}) - v(S \cup \{j\}) + v(S)), \quad (3.2)$$

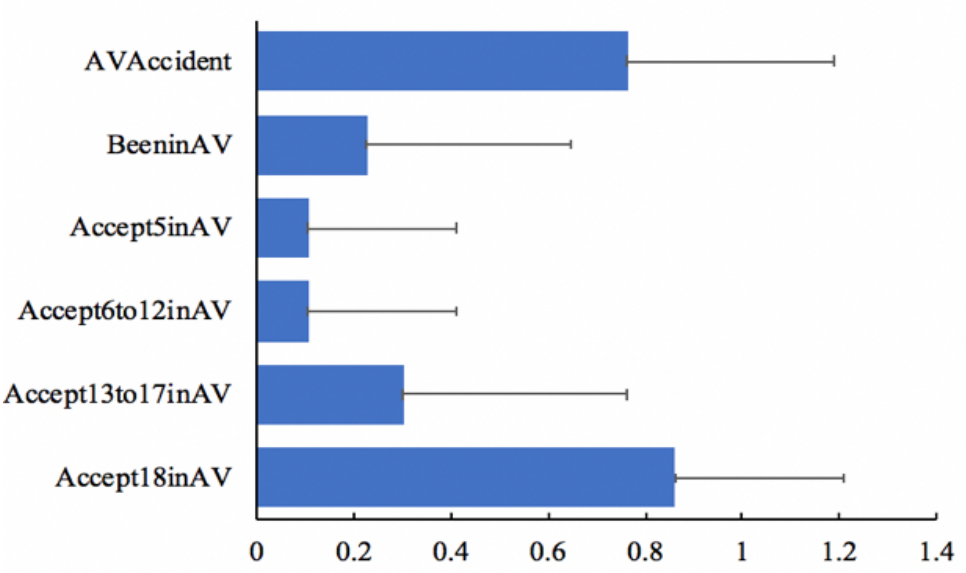
Thus, SHAP can produce an *n* by *n* interaction matrix and automatically can identify the strongest interaction effect given one specific feature. In this research, after training the XGBoost model, SHAP was used to explain the model predictions (Lundberg and Lee, 2017) by calculating the importance of each feature, by evaluating the interaction effects between the features globally, and by explaining individual predictions locally.

3.3 Results

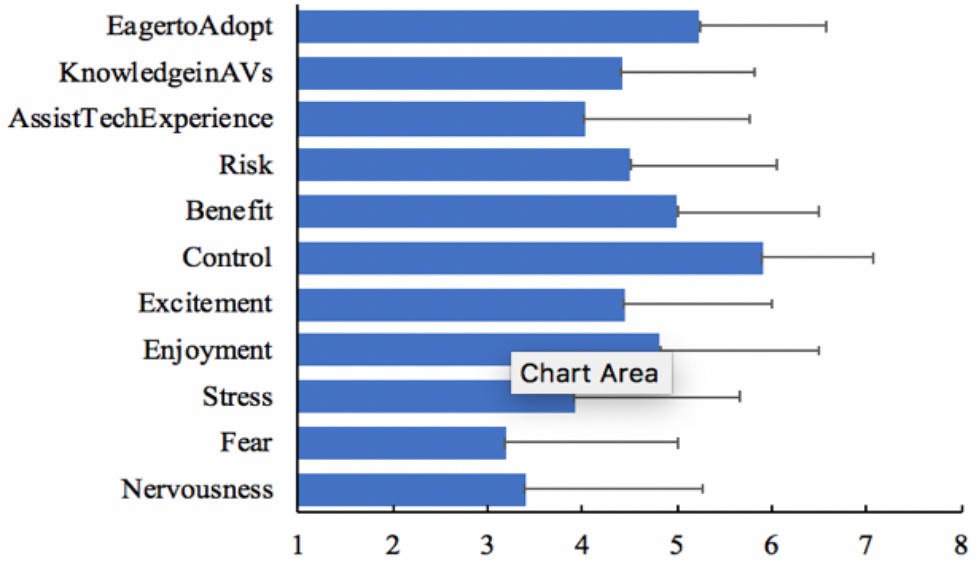
We calculated participants' mean responses and the standard deviations as shown in Figure 3.2 . The knowledge-related questions indicated that the majority of the participants had a relatively high level of knowledge about AVs — 75.1% had a high level of eagerness to adopt a new technology (i.e., by high, we mean a Likert scale value greater than or equal to 5, moderate refers to a Likert scale value of 4, and low refers to a Likert scale value less than or equal to 3),

51% had a high level of knowledge in AVs, and 76.4% of the participants knew about accidents related to AVs. As for the experience related questions, the majority showed a low level of experience in AVs—46% of the participants had a high level of experience in ADAS and 77.3% had never been in an AV. Furthermore, the majority considered AVs as beneficial (71%), but risky (57%). In regard to behavioral assessment of AVs, 89% of the participants were reluctant to let a child under 5 or between 6 and 12 use an AV alone and 70% were reluctant to let a child between 13 and 17 use an AV alone. However, 86% were willing to let a child above 18 use an AV alone. Feelings related questions showed that the majority of the participants reported a high level of control (91%) and a high level of excitement (51%) and enjoyment (64%) while driving. In addition, 58% of the participants had a low level of fear and nervousness of driving, but 44% of the participants considered driving as being stressful.

The performance of the XGBoost prediction model, including accuracy, ROC_AUC, precision, recall, and F1 measure, is shown in Table 3.3 using a 10-fold cross-validation strategy. To compare the performance of XGBoost with other algorithms (see Table 3.3), we also performed a 10-fold cross- validation strategy. We found that XGBoost performed the best across almost all the metrics (except precision) among the list of the machine learning models including logistic regression, decision trees, naive Bayes, linear SVM, and random forest.



(a)



(b)

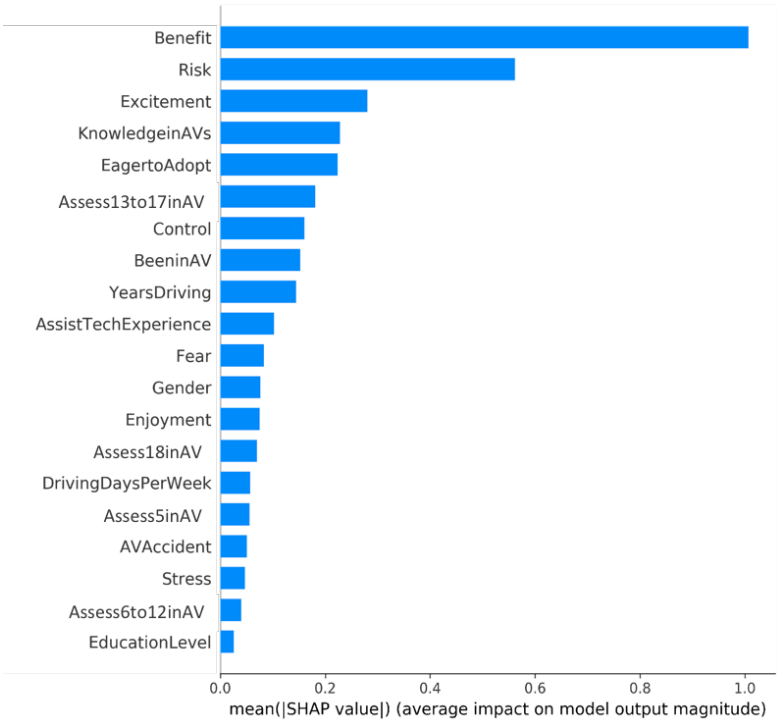
Figure 3.2: Mean values and standard deviations of the predictor variables. (a) "0" = No, 1 = "Yes"; (b) "1" = Extremely low, "2" = Moderately low, "3" = Slightly low, "4" = Neither low nor high, "5" = Slightly high, "6" = Moderately high, "7" = Extremely high.

Table 3.3:Performance measures comparison between different models.

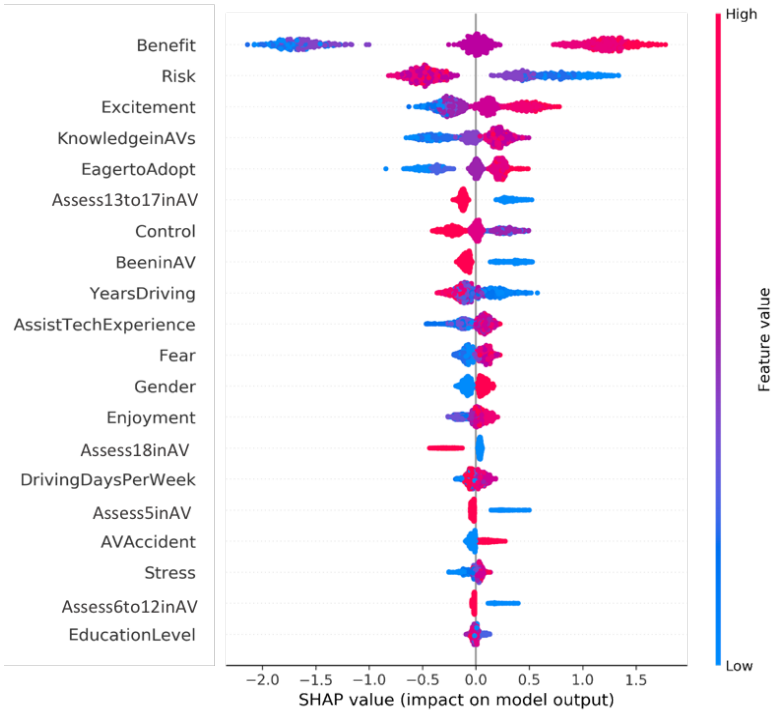
Models	Accuracy	ROC_AUC	Precision	Recall	F1 Measure
Logistic Regression	83.1%	0.90	82.1%	82.6%	82.3%
Decision Tree	83.5%	0.87	**82.9%**	82.8%	82.9%
Naïve Bayes	81.6%	0.90	81.2%	80.8%	81.0%
Linear SVM	84.4%	0.91	82.8%	84.3%	83.5%
Random Forest	83.1%	0.90	81.3%	83.3%	82.3%
XGBoost	**85.5%**	**0.92**	82.5%	**91.6%**	**86.8%**

3.3.1 Importance of predictor variables

To understand the importance of each factor in predicting perceived trust in AVs, we examined SHAP feature (i.e., predictor variable) importance and summary plots. The SHAP feature importance plot sorts the features by the mean of the absolute SHAP value over all the samples i.e., $\frac{1}{M}\sum_{j=1}^{M}|\phi_{ij}(v)|$, where M is the total number of the samples. The SHAP summary plot also combines feature importance with feature effect. Note the unit of the SHAP value here is log odds as the objective function was set as logistic regression in training the XGBoost model. The summary plot lists the most significant factors in a descending order as illustrated in Figure 3.3(a). The top factors (e.g., Benefit, Risk, Excitement, KnowledgeinAVs, EagertoAdopt) contributed more to the prediction. To obtain more information about the factors, we also explored the summary plot in Figure 3.3(b).



(a)



(b)

Figure 3.3: (a) SHAP feature importance plots (b) SHAP summary plot.

Each data point (i.e., each participant) has three characteristics, including 1) the vertical location that shows importance ranking based on the overall SHAP value of a particular predictor factor, 2) the horizontal spread that depicts whether the value has a small or large effect on the prediction, and 3) the color coding that describes the value of the factor from low (i.e., blue) to high (i.e., red) gradually. For instance, a small value of the Benefit factor has shown to reduce the log odds of the prediction of trust by almost 2.5, whereas a large value of the Benefit factor increases the prediction by almost 2. Such results not only show the importance of the predictor variables, but also help us understand how they influence the prediction results. Furthermore, the spread of the important factors tends to be wider than those of the unimportant factors, and the SHAP value of the majority of the unimportant factors tends to be around 0, such as EducationLevel.

3.3.2 Dependence plot

To further understand the relationship between the predictor variables and the response variable, we examined their individual SHAP dependence plots which can capture both, the main effects of individual predictor variables and the interaction effects between predictor variables. Figure 5 shows the SHAP dependence plots of the top five most important factors (i.e., Benefit, Risk, Excitement, KnowledgeinAVs, and EagertoAdopt) and a continuous variable, i.e., YearsDriving. For instance, to understand the impact of Benefit on trust as captured by the XGBoost model, the SHAP dependence plot is shown in Figure 3.4(a). The horizontal axis represents the actual values of the Benefit factor from the dataset, and the vertical axis represents the effect of the factor on the prediction. For the main effect, the plot shows an increasing trend between the factor Benefit and the target trust. It also shows the interaction effect between Benefit and BeeninAV automatically selected by the SHAP model. Out of the participants who scored low

for perceived benefits of AVs, those who had experience with AVs trusted AVs more than those who had no experience. On the other hand, out of the participants who scored high for perceived benefits of AVs, those who had experience with AVs trusted AVs less than those who had no experience with AVs. The SHAP dependence plot of Risk is illustrated in Figure 3.4(b). We can observe that risk is negatively correlated with trust in AVs. Meanwhile, out of the participants who scored low for risks in using AVs, those who had no experience in AVs trusted AVs more than those with experience. On the other hand, out of the participants who scored high for risk in AVs, those who had experience with AVs trusted AVs more than those who did not. The effect of Excitement on trust is illustrated in Figure 3.4(c). The higher the excitement about manual driving, the higher the likelihood to trust AVs. And among the participants with a low level of excitement about driving, those who scored high for perceived risks in AVs trusted AVs less than those who scored low for perceived risks. However, among the participants with a high level of excitement about driving, those who scored high for perceived risks in AVs trusted AVs more than those who scored low for perceived risks in AVs. Figure 3.4(d) illustrates the effect of KnowledgeinAVs on trust. The increasing slope indicates that the more the Knowledge in AVs, the higher the likelihood to trust AVs. For the participants who rated low in knowledge in AVs, those with low perceived risks in AVs trusted AVs more than those with high perceived risks in AVs. However, when the participants rated high in knowledge in AVs, those with high perceived risks in AVs trusted AVs more than those with low perceived risks in AVs. The increasing slope in Figure 3.4(e) shows that the more eager the participants are to adopt a new technology, the higher the likelihood is to trust AVs. Out of the participants who were not eager to adopt a new technology, the interaction effect was not clear. However, out of the participants who were eager to adopt a new technology, those being not fearful of driving trusted AVs more than those being fearful of driving. In Figure 3.4(f),

we see a decreasing slope which illustrates that people with more experience in driving are less likely to trust AVs. For the participants with driving experience between 10 and 40 years, those who reported a high level of perceived benefits trusted AVs more than those who reported a low level of perceived benefits.

3.3.3 Main effects and interaction effects

The SHAP dependence plot has rich information, which incorporates both main effects of individual predictor variables and interaction effects between two predictor variables. The interaction effects are demonstrated by the vertical dispersion as shown in Figure 3.4.

Such interaction shows the effect of the two predictor variables on the response variable at the same time. We can also separate the main effects and interaction effects in individual plots. Take the Risk SHAP dependence plot in Figure 3.5(b) as an example. Its main effect and interaction effect with BeeninAV are shown in Figure 3.5(a) and Figure 3.5(b). There is little vertical dispersion in the main effect. The interaction effect is also more apparent suggesting that at lower Risk levels, participants who experienced AVs trusted AVs less than those who did not experience AVs. However, at higher Risk levels, participants who experienced AVs trusted AVs more than those who did not experience AVs. Take the YearsDriving as another example. Its main effect and interaction effect with Benefit are shown in Figure 3.5(c) and Figure 3.5(d). Also, less vertical dispersion is observed in the main effect plot, and the interaction effect tends to be more apparent. That is, only when YearsDriving is larger than 10 and smaller than 40, more Benefits lead to a stronger likelihood to trust AVs.

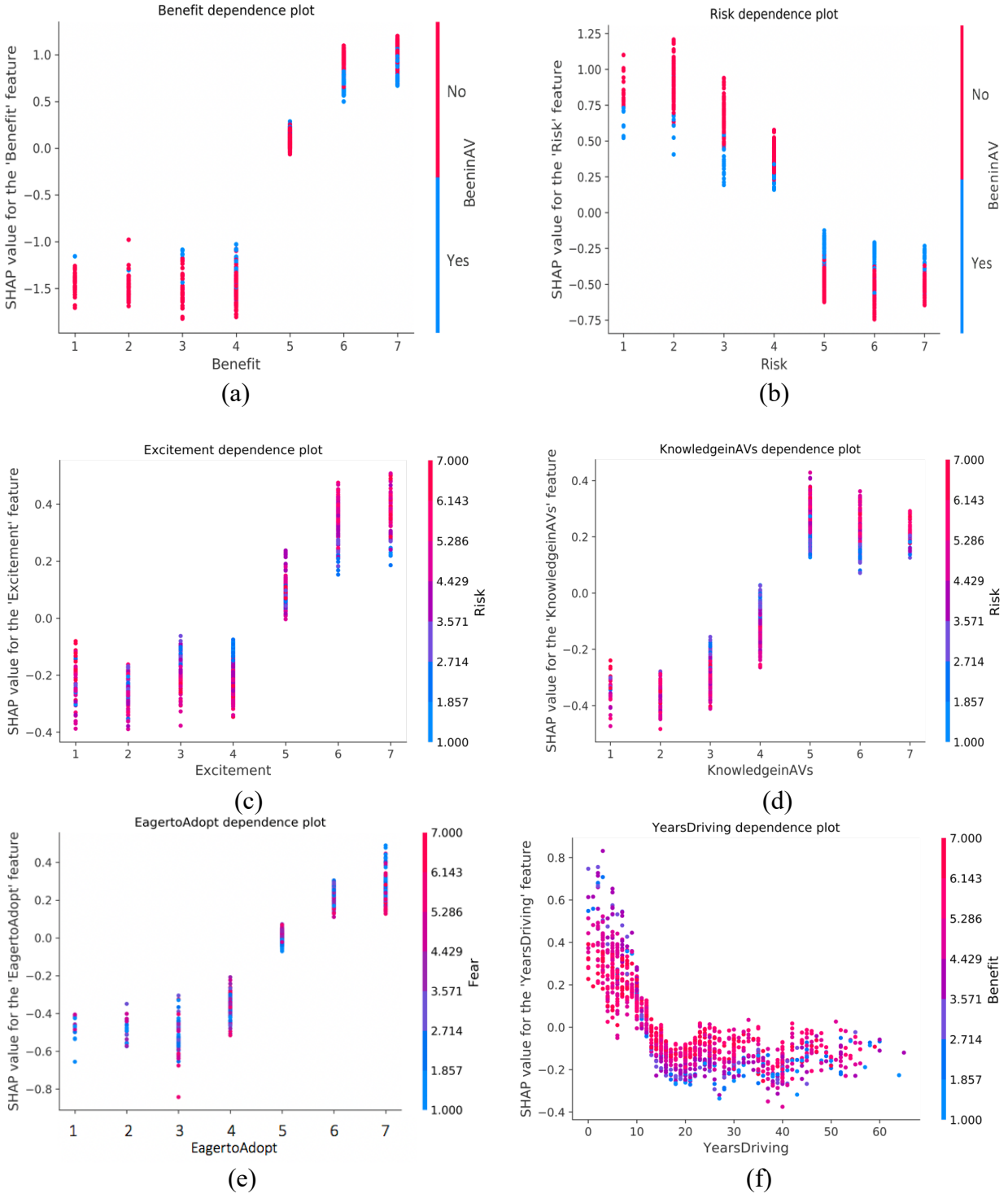


Figure 3.4: SHAP dependence plots. (a) Benefits, (b) Risk, (c) Excitement, (d) KnowledgeinAVs, (e) EagertoAdopt, and (f) YearsDriving. “1” = Extremely low, “2” = Moderately low, “3” = Slightly low, “4” = Neither low nor high, “5” = Slightly high, “6” = Moderately high, “7” = Extremely high.

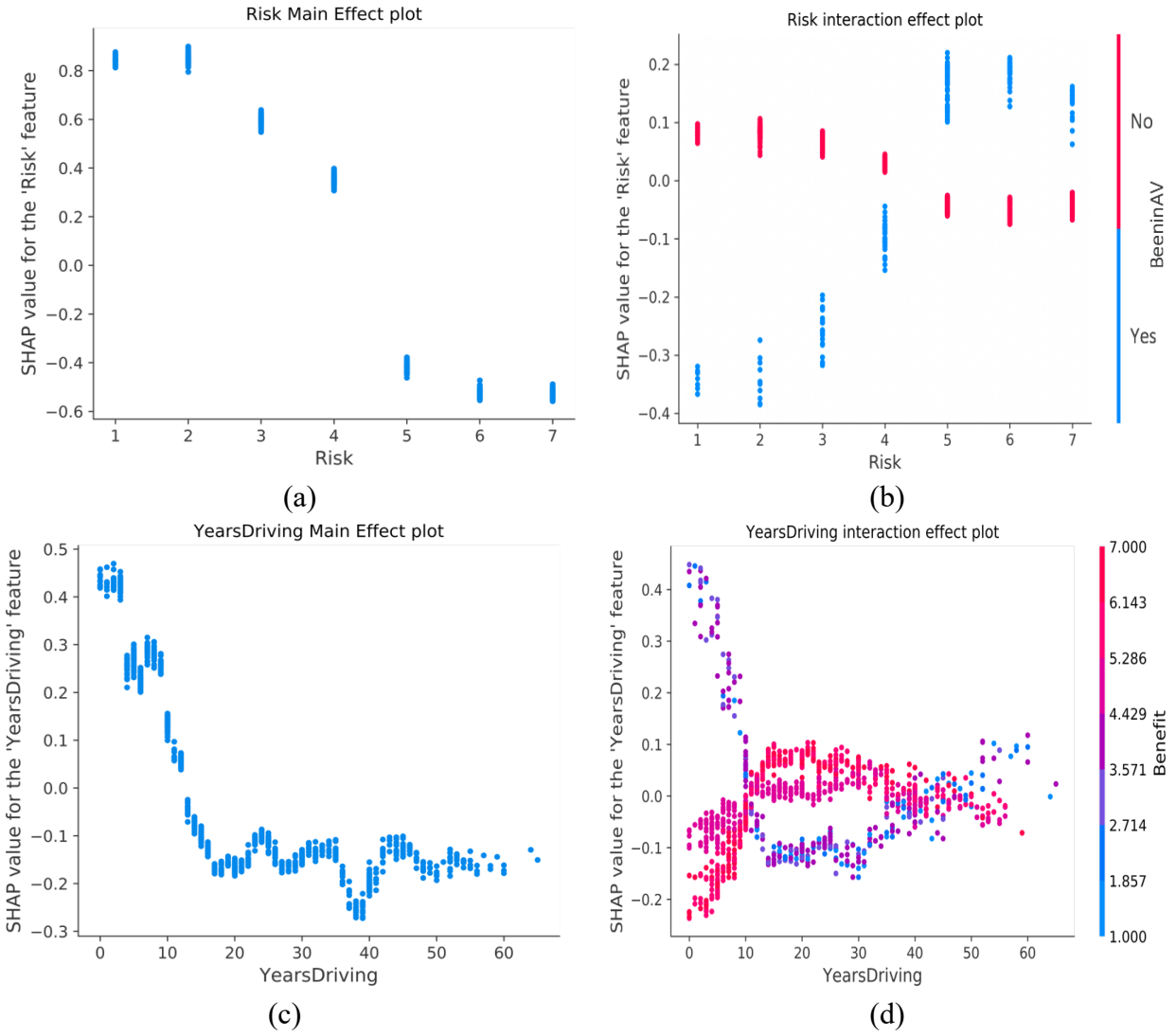
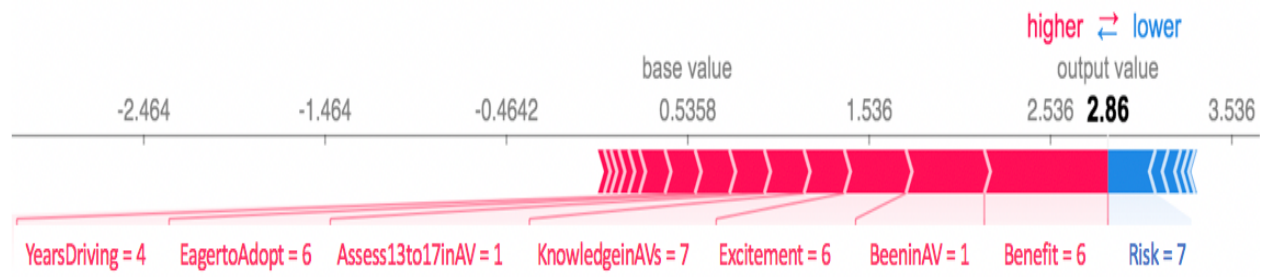


Figure 3.5: SHAP main effects and interaction effects derived from SHAP dependence plots. "1" = Extremely low, "2" = Moderately low, "3" = Slightly low, "4" = Neither low nor high, "5" = Slightly high, "6" = Moderately high, "7" = Extremely high.

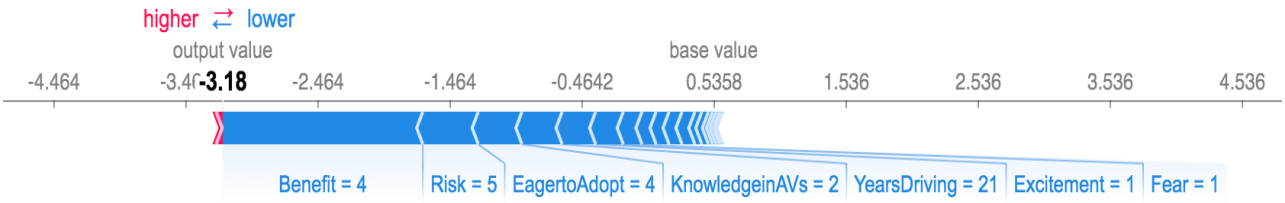
3.3.4 SHAP local explanation

In order to show how SHAP explains individual cases, we tested it on two randomly selected observations as illustrated in Figure 3.6. The plots show the different factors contributing to pushing the output value from the base value which represents the average model output over the training dataset. The base value is defined as the mean prediction value, which is 0.5358 in our

case (Lundberg et al., 2018). Factors pushing the SHAP value (i.e., log odds) to be larger are shown in red while those pushing the SHAP value to be smaller are shown in blue. In Figure 3.6(a), the model produced a large SHAP value in predicting trust which was consistent with the ground truth (i.e., trust) because the participant perceived the AV with a high level of Benefits (i.e., 6), BeeninAV = Yes, a high level of Excitement (i.e., 6), a high level of KnowledgeinAVs (i.e., 7), Assess13to17inAV = Yes, a high level of EagertoAdopt (i.e., 6), YearsDriving (i.e., 4), even though the participant perceived the AV with a high level of Risk (i.e., 7). In Figure 3.6(b), the model produced a small SHAP value, which was consistent with the ground truth (i.e., distrust) mainly due to a neutral level of Benefit, a high level of Risk (i.e., 5), a neutral level of EagertoAdopt (i.e., 4), a low level of KnowledgeinAVs (i.e., 2), 21 YearsDriving, a low level of Excitement (i.e., 1), and a low level of Fear (i.e., 1).



(a)



(b)

Figure 3.6: SHAP individual explanations of trust prediction for randomly selected participants with (a) ground truth = trust and (b) ground truth = distrust.

3.4 Discussion

3.4.1 Predictability and explainability

XGBoost is an efficient and easy to use algorithm for tabular data classification which delivers high performance and accuracy as compared to other algorithms (Chen & Guestrin, 2016). In this research, we used XGBoost to predict people's trust in AVs with satisfactory performance. Compared to other machine learning models, XGBoost performed the best among various metrics, including accuracy, ROC_AUC, recall, and F1 measure (see Table 3.3). The model converged within 60 iterations in our experiment and proved to be a feasible solution to predict trust in AVs.

In order to improve the explainability of the XGBoost model, we used a SHAP explainer which offers a high level of model interpretability (Lundberg & Lee, 2017). SHAP has a fast implementation for tree-based models (e.g., XGBoost) which overcomes the biggest barrier (i.e., slow computation) for adoption of Shapley values. On top of the advantage of fast implementation, SHAP provides two more advantages including global and local interpretability. The global interpretability is represented by the contribution of the SHAP values in the model predictive decision. It can represent the negative and positive effect of the most important factors on the model prediction as shown in Figure 3.3. Such global interpretability tends to be similar to the feature effect plot in linear regression models. Furthermore, the model is able to show interaction effects between different predictor variables indicating how they influence the prediction results as evidenced in Figure 3.4 and Figure 3.5. As for the local interpretability, SHAP enables us to explain the prediction of each observation since each one gets its own set of SHAP values as illustrated in Figure 3.6. With the local and global interpretability comes the power of SHAP in providing a high level of model explainability.

3.4.2 Important factors in predicting trust

Compared to linear regression models, our method uncovered the factor importance in predicting trust using the SHAP feature importance plots and the SHAP summary plot as shown in Figure 3.3. Among all the predictor variables, the Benefit factor ranked the most important and was positively correlated with trust, consistent with previous research (Choi & Ji, 2015; Bearth & Siegrist, 2016). Furthermore, we also found an interaction effect between Benefit and BeeninAV (see Figure 3.4(a)). Even when the participants perceived the AVs with low benefits, their interaction with AVs could potentially improve their trust in them. This was consistent with Brell et al. (2019), which showed that the experience with AVs significantly increased the perception of the benefits in AVs.

The second most important factor was risk (Figure 3.3). In line with prior studies (Numan, 1998; Kim et al., 2008; Pavlou, 2003), our results showed that an increase in risk led to a decrease in trust. Risk was found to interact with BeeninAV (Figure 3.4(b)). When the participants viewed AVs to be risky, experience with AV could potentially improve their trust in AV. This was also in concordance to previous research (Brell et al., 2019), which showed a decrease in risk perception in AVs with the increase of experience in AVs. Therefore, it is important that automotive manufacturers give more chances for the public (especially for those who do not perceive AVs with benefits and/or high risks) to test AVs in order to improve their trust in AVs.

While both the third and fourth most important factors, i.e., Excitement and KnowledgeinAVs were positively correlated with trust in AVs. Risk was found to interact with Excitement (Figure 3.4(c)) and KnowledgeinAVs (Figure 3.4(d)). When the participants were not very excited about manual driving, they tended to trust the AVs more if the risk was low. Silberg et al. (2013) found that people who were less passionate about driving were more likely to lean

toward using AVs if it was safe. When the participants were excited about manual driving, they trusted the AV more even if the risk was higher. Such trust, however, could be overtrust associated with strong emotions such as excitement. For example, Dingus et al. (2016) argued that excited or angry drivers were more likely to take risky driving even in highly automated driving. An increase in KnowledgeinAVs increased the trust in AVs (Figure 3.4(d)) which was consistent with previous studies such as (Khastgir et al., 2018). When the participants knew more about AVs, they still trusted AVs even with a likely high level of risks. This might be explained that the degree of knowledge about risks affected the perception of the risks of AVs and trust in AVs. For instance, the more one knows about the risk, the higher the chances to accept it (Schmidt, 2004).

The EagertoAdopt factor was ranked number 5, and an increase in eagerness to adopt a technology increased the chances of trusting the AV which was in line with previous research (Edmonds, 2019; Raue et al., 2019) (see Figure 3.4(e)). We also found that Fear affected the impact of EagertoAdopt on trust—at a high level of eagerness to adopt a new technology, a low fear of manual driving increased the chances of trusting the AV. Fear, which is an important factor in technology adoption, was shown to shape judgements, choices, and perception of risks (Lerner & Keltner, 2001). According to Shoemaker (2018), fearless driving was associated with no fear of change, thus leading to an eagerness of technology adoption. Other factors involved in the study were less important compared to the ones listed above. Although Assess13to17inAV was ranked number 6, it was surprising to see that Assess5inAV and Assess6to12inAV were less important in predicting trust in AVs. Intuitive, without trust in AVs, a parent would not let children be in AVs. However, in our survey, we did not specify if they were the participants' children. Further research is needed to address this issue. Gender, age (years of driving), and education level were also found to be less important. However, as seen in Figure 3.4(f), we found that trust was shown to decrease

with an increase in the number of driving years. Furthermore, Benefit affected the impact of DrivingYears on trust—for larger than 10 years and smaller than 40 years of driving experience, high benefit increased the trust in AVs. In line with previous research, old people showed more concerns about trusting AVs despite its benefits in maintaining their mobility (Schoettle & Sivak, 2016) while young drivers with less experience tended to be risky drivers.

As a summary, the measured trust is based on dispositional trust and initial learned trust (see Hoff and Bashir, 2015). The dispositional trust shows participants' overall tendency without any context of AVs and the initial learned trust is dependent on their previous knowledge or past experience (e.g., news reports on AV accidents) prior to interacting with AVs. This is because the majority of the participants (i.e., 77.3*%*) had no chance to interact with AVs and there was no interaction between the participants and AVs during this study. However, the dispositional trust and the initial learned trust measured in our paper are the baseline to form people's trust in AVs. Prior to any interaction with AVs, people have an inherent level of dispositional trust which is one of the major factors that influences people's purchase or use of AVs. Individual differences, such as age, gender, educational levels, as well as their learned knowledge about and experience in AVs shaped their perceived risks in and benefits of AVs, which in terms influence their dispositional and initial learned trust. Between these two types of trust measured in the survey, we found that the variables related to dispositional trust were more important and predictive than those related to initial learned trust as shown in Figure 3.3a. Nevertheless, unlike previous studies, the most important contribution of this study was proposing a trust prediction model with explainability to understand participants' trust in AVs. Automotive manufacturers can potentially make use of the relationships between these important factors and their trust to improve acceptance and adoption

of AVs by providing training, spreading the benefits of AVs, explaining the possible risks, improving the design of the system, and creating appropriate emotional responses to AVs.

3.5 Limitations

First, due to the cross-sectional study design, we cannot examine how people's opinions and judgments change over time or in response to new information about self-driving vehicles or experiences with advanced vehicle technologies. Longitudinal studies may be needed in order to understand the dynamic trust relationships between users and AVs that start long before their first contact with the system (Ekman et al., 2018). Second, it was difficult for us to make sure the superior quality of the survey data from AMT. In this research, we made use of various techniques to combat that, including shorter surveys, removing invalid data by examining their survey completion time and data patterns. However, quality can be affected by the compensation rate (Buhrmester et al., 2011) and running the screening procedures mentioned above might be not enough to ensure a high quality of responses. Moreover, the participants' age distribution was not equally distributed. It is necessary to control the age factors as it can potentially be an important factor in trust prediction. Third, our survey was quantitative without any qualitative data to explain our prediction model. It would be also important to verify such explanations using qualitative data from the participants themselves with open-ended questions. Finally, judging whether one would use a self-driving car without ever having seen one or experienced riding in one is a difficult task. Although a previous study (Raue et al., 2019) showed that peoples' experience in manual driving and ADAS affects their feelings and perception of automated driving, it would be difficult to shape exactly how they would feel about AVs. Thus, an optimal assessment of participants' trust and feelings about AVs would be to experience it in automated driving (Ruijten et al. 2018).

3.6 Conclusion

In this study, we attempted to predict trust in AVs with both accuracy and explainability using XGBoost and SHAP models. To predict trust in AVs, we conducted an online survey to collect various variables that were related to participants' trust in AVs. The survey data were then used to train and test the XGBoost model. To help better understand the XGBoost model, SHAP was used to explain the trust predictions by identifying the most important predictor variables, by examining their interaction effects, and by illustrating individual explanation cases. Compared with previous trust predictions studies, our proposed method combines the benefits of XGBoost and SHAP with good explainability and predictability of the trust model.

CHAPTER 4

Predicting Drivers' Situational Trust Using Physiological Measurements in Conditional AVs

4.1 Introduction

The second phase of the proposed research aims to predict drivers' trust in conditional AVs using physiological measurements in real-time. In chapters 3 and 5, we used subjective measures such as self-reported measures to assess trust. Previous researchers argued that self-reports cannot capture real-time changes in trust specifically during real-world driving since it is not practical to ask the participants to fill the same questionnaire multiple times during the experiment and their answers can sometimes be biased (Hergeth et al., 2016; Walker et al., 2019). Therefore, to objectively measure trust in real-time psychophysiological measurements such as galvanic skin response (GSR), gaze behavior, electrocardiogram (ECG), and electroencephalography (EEG) are needed (Hergeth et al., 2016; Walker et al., 2019; Akash et al., 2018). These measurements can provide a direct indicator of drivers' trust state in real-time. Accordingly, there is an urgent need to develop prediction models to infer automation trust in real-time for different types of drivers to successfully interact with AVs. Not meeting this need represents an important issue in automated driving because, without measuring trust in real-time, both misuse and disuse will continue to be

existing in automated driving. Thus, the overall objective of this study is to predict situational trust in real time in conditional AVs.

4.2 Method

4.2.1 Participants

A total number of 74 university students were recruited for this study. Due to malfunction of physiological sensors and driving simulator, 15 participants were excluded and 59 participants (as shown in Table 4.1) (mean age = 21.3; standard deviation = 2.9; range = 18-33; 26 females and 33 males) were included for further analysis. All the participants had a valid driver's license and had normal or corrected-to-normal vision. Participants received $25 in compensation for about an hour of participation. The study was approved by the Institutional Review Board at the University of Michigan.

Table 4.1: Distribution of participants under the different conditions.

Control		FA		Misses	
Urban	Suburban	Urban	Suburban	Urban	Suburban
16	16	22	22	21	21

4.2.2 Apparatus and stimuli

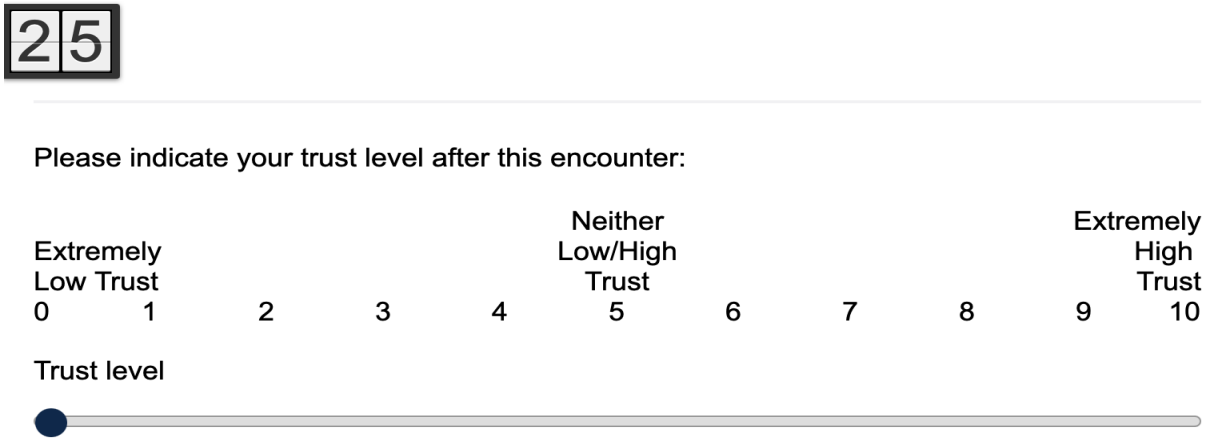
The study was conducted in a desktop-based driving simulator from Realtime Technologies Inc. (RTI, MI, USA) to create a driving experience close to a real car. As shown in Figure 4.1a, the simulator setup was composed of three LCD monitors integrated with a Logitech driving kit. Two other touchscreens (i.e., tablet and phone) were positioned to the right side of the participants for the NDRT and trust rating questions. The NDRT in this study was a Tetris game coded using

PyGame library in Python. The game was designed in a way that the participants need to drag the blocks otherwise they don't move. With this design, participants can leave the game to handle takeover scenarios and get back to where they were before the TOR. A questionnaire to evaluate participants' trust was designed in Qualtrics (Provo, UT, www.qualtrics.com), web-based software to create surveys. To evaluate participants' trust, the questionnaire was designed with one question that was prompted on participants' screen every 25 seconds asking them to rate their trust on a scale from 0 to 10 (see Figure 4.1b).

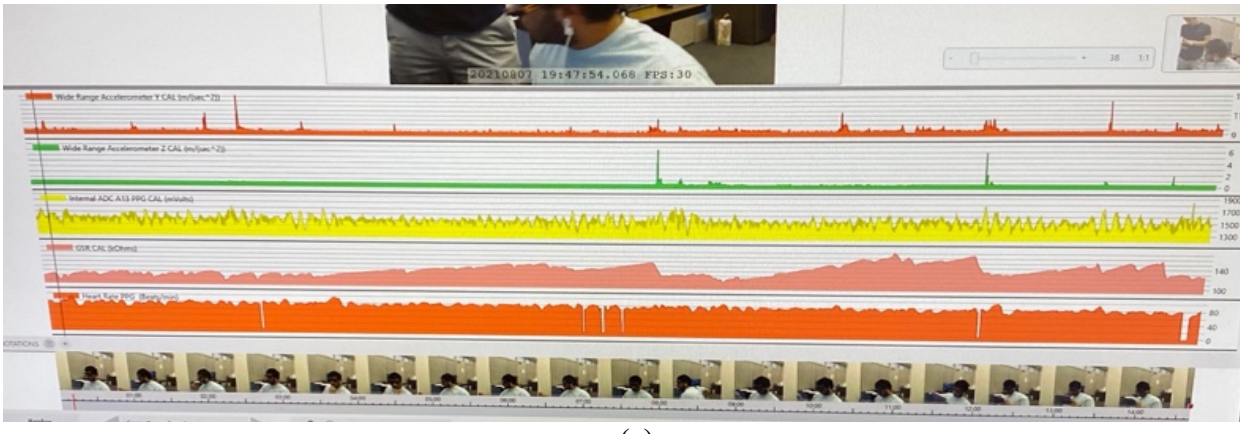
Participants were able to control the simulated vehicle using a steering wheel and the pedal system. To engage the autonomous mode, participants needed to press a red button on the steering wheel. Once the autonomous mode is engaged, participants will hear an auditory warning saying, "Automated mode engaged". The driving simulation was programmed to simulate SAE Level 3 automation and participants were able to control the vehicle laterally and longitudinally. Whenever a takeover is needed, participants will hear an audio warning ("Takeover") and the automated mode will be automatically disactivated for drivers to take control. If the drivers don't take control within the time limit, an emergency stop ("Emergency Stop") audio will be issued to avoid any crash.

Pupil Lab's Pupil Core eye tracker headset was used to measure participants' gaze positional data. The sampling rate of the eye-tracking system is around 15 Hz. The Shimmer3 GSR+ Unit (Shimmer, MA, USA) was used to measure the skin conductance using electrodes on the foot arch and to capture an Optical Pulse/ photoplethysmogram (PPG) signal using the Shimmer ear clip and converting it to a heart rate (HR). The data from Shimmer was collected at a sampling rate of 128 Hz. The iMotions software (iMotions, MA, USA) was used for physiological data synchronization in real-time (see Figure 4.1c).

(a)



(b)



(c)

Figure 4.1: (a) Experiment setup (b) Trust change self-report question (c) iMotion software.

4.2.3 Experimental design

The trust prediction experiment is a between-subject design, in which the participants will be randomly assigned to three conditions. It is argued that automation trust should be examined both in normal operating conditions and in emergency situations where the system encounters limitations (Madhani et al. 2002). In this study, participants were required to take over from automated driving in emergency situations. Eight takeover scenarios were designed for this study, where four takeover scenarios happened in rural areas and the other four happened in urban areas and the order of urban and rural takeover scenarios was counterbalanced. Typical roadway features were used when a takeover was needed (i.e., (1) deers ahead, (2) bicyclist crossing ahead (3) construction zone ahead (4) vehicle sudden stop ahead (5) pedestrians crossing ahead (6) bus sudden stop ahead (7) construction zone ahead (8) police vehicle on shoulder) (See Figure 4.2 and Figure 4.3). The AV error types correspond to three conditions: 1) control condition, i.e., all the eight TORs are true alarms, 2) false alarms, i.e., of all the eight TORs, the 2nd, 3rd, 5th, and 6th are false alarms, and 3) misses, i.e., of all the eight TORs, the 2nd, 3rd, 5th, and 6th are misses. The purpose of this design was to elicit different levels of trust since it was shown that both misses and false alarms degraded operator trust in automation (Pop et al., 2015).

4.2.4 Experimental procedure

Upon arrival, participants were asked to complete a consent form as well as an online demographic survey. After the survey, we explained the experiment to the participants' and showed them a short video regarding the tasks they need to do. Participants then completed a training session to familiarize them with the driving simulator and the experiment flow. Participants were informed that the car will be able to take situational decisions, but the driver

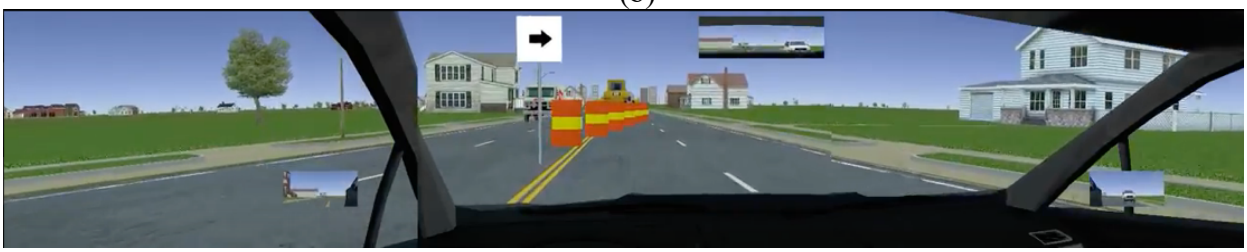
must be alert and ready to takeover whenever it is needed. We further explained that the AV can fail to detect some obstacles for the participants going through the misses condition and that the AV can give false alarms for takeovers for those going through the FA condition. Next, we calibrated the eye-tracker device by asking the participant to look at ten targets on the front screens. After, we attached the GSR electrode to the left foot of the participants and the PPG probe to their left ear lobe. Each drive (i.e., urban or suburban) took around 15 minutes and the whole experiment lasted around 75 minutes. In order to create the ground truth of trust prediction model in real time, the participants were asked to respond to a single-item trust prompt on a scale from 0 to 10, "Please indicate your trust level after this encounter" (Hergeth et al., 2016). Following Desai et al. (2013), participants will be prompted for this trust measure every 25 seconds to ensure that they are not overwhelmed.

4.2.5 Comparison between the three tested conditions

Results showed that there was a statistically significant difference in trust between the three tested conditions (i.e., control, misses, and FA) as determined by one-way ANOVA $F(2,47) = 22.323$, $p < 0.001$. A Tukey post hoc test revealed that trust was significantly higher in the control (mean = 7.967) and FA (mean = 7.699) conditions compared to the misses condition (mean =5.597) ($p < 0.001$). There was no significant difference between the FA and the control condition. Since the FA condition didn't reduce participants' trust, we did not include it in the trust prediction model. Figure 4.4 shows average participants' trust with respect to the rating order for the three tested conditions (i.e., control, misses, and FA). We didn't find any significant difference in the seven important features between the three tested conditions (i.e., control, misses, and FA) as shown in Table 4.2.

(a)

(b)



(c)

(d)

Figure 4.2: Takeover events in suburban areas (a) deers ahead (b) bicyclist crossing ahead (c) construction zone ahead (d) vehicle sudden stop ahead.

(a)

(b)

(c)

(d)

Figure 4.3: Takeover events in urban areas (a) pedestrians crossing ahead (b) bus sudden stop ahead (c) construction zone ahead (d) police vehicle on shoulder.

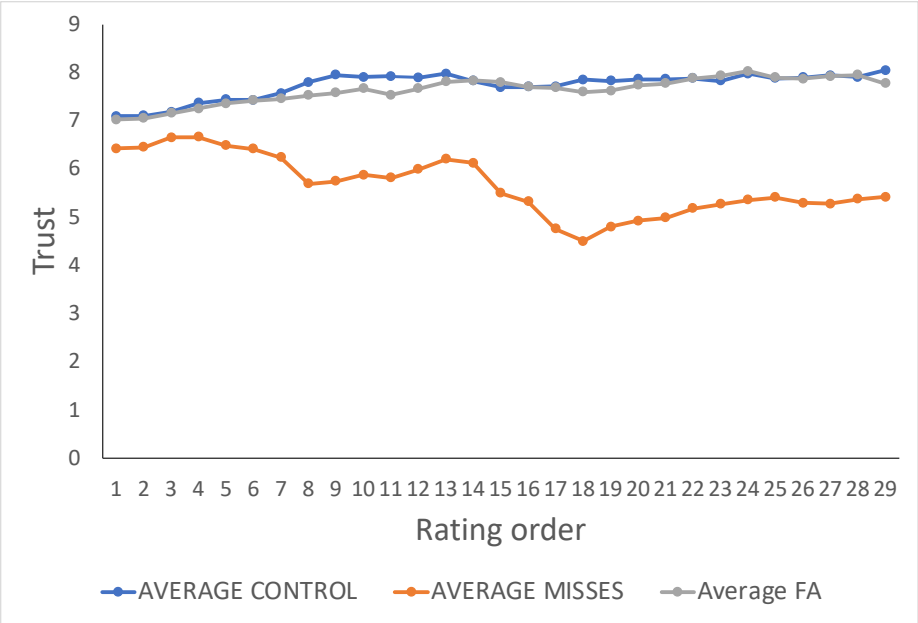


Figure 4.4: Comparison of the average participants' trust with respect to the rating order for the control, misses, and FA conditions.

Table 4.2: Statistical analysis results for the seven important features. Kruskal Wallis test was used when the normality assumption was violated.

Feature	Condition	Mean ± SD	Test	*P*
Mean_HR_max	Control Misses False Alarm	100.67 ± 16.11 93.99 ± 20.15 96.17 ± 10.14	One-way ANOVA	0.526
Mean_HRV	Control Misses False Alarm	70.28 ± 11.24 75.57 ± 17.18 72.36 ± 15.42	Kruskal Wallis	0.484
Number_of_fixations_center	Control Misses False Alarm	54.64 ± 33.68 65.75 ± 35.94 43.41 ± 22.33	One-way ANOVA	0.090
Mean_GSR	Control Misses False Alarm	0.59 ± 1.01 0.18 ± 0.14 0.11 ± 0.13	Kruskal Wallis	0.208
Number_of_fixations_tablet	Control Misses False Alarm	182.30 ± 81.62 127.07 ± 81.52 148.69 ± 85.76	One-way ANOVA	0.216
Mean_dispersion_tablet	Control Misses False Alarm	0.91 ± 0.18 0.82 ± 0.30 0.94 ± 0.22	One-way ANOVA	0.376
Mean_duration_tablet	Control Misses False Alarm	128.18 ± 30.61 111.98 ± 43.29 130.57 ± 36.59	One-way ANOVA	0.290

4.3 Trust prediction model development

Physiological (i.e., galvanic skin response, heart rate, and eye-tracking metrics) and self-reported data were collected for the trust model development. To train the trust prediction model, 17 features were extracted from the data (see Table 4.3). A 5-fold cross validation was used to optimize the F1-score of the prediction using a randomized search for hyperparameters.

4.3.1 Data pre-processing:

The GSR is composed of phasic (i.e., fast variation of skin conductance) and tonic (slow variation of skin conductance) phases. In this study, the phasic component was used since it captures the GSR changes in seconds. Therefore, we used a continuous decomposition technique (i.e., Ledalab in MATLAB) (Benedek and Kaernbach 2010). The iMotion software was used to extract the heart rate related measures from the RR interval. For eye-tracking data, Pupil Player software was used for exporting the data collected in Pupil Core for further analysis.

4.3.2 Model features

As shown in the previous sections, our data was collected using different sensors (i.e., Shimmer and eye-tracking) and systems (i.e., Qualtrics). Therefore, we synchronized the time between GSR, HR, eye-tracking, and continuous trust data using timestamp. After time synchronization, we used a sliding time window of 25 seconds to extract a series of GSR, HR, and eye-tracking values within that window. Therefore, each self-reported rating was associated with a series of GSR, HR, and eye-tracking values at the same timestamp. The extracted 17 features are listed in Table 4.3.

Table 4.3: Description of the generated features.

Physiological Data	Model Features
Heart rate (HR)	Heart rate max, heart rate variability, inter-beat interval
Fixation	Number of fixations on the center, left, right, and NDRT screens
Duration	Duration of fixations on the center, left, right, and NDRT screens
Dispersion	Distance between all gaze locations during a fixation on the center, left, right, and NDRT screens
Galvanic skin response (GSR)	Mean and max of galvanic skin response in phasic phase

4.3.3 Model development

The trust prediction model was trained with an XGBoost model (Chen and Guestrin, 2016) for the following reasons. First, XGBoost is a decision tree model that combines multiple trees where each decision tree learns from the previous one to build a robust model. Second, XGBoost has the power to perform parallel processing and the advantage to improve the learning process without overfitting. In addition, XGBoost works with missing data without affecting the model performance. Third, XGBoost model can provide feature importance and helps in interpreting the model predictions with the usage of SHAP explainer. The XGBoost model performance was also compared with other algorithms (e.g., logistic regression (LR), decision tree (DT), Naïve Bayes (NB), and K-nearest neighbors (KNN)). The response variable was defined as a binary one (i.e., trust = 1 (trust value > 5, sample size = 1745) and distrust = 1 (trust value < 5), sample size = 484). The objective function used is binary: logistic. The performance metrics used to evaluate the XGBoost model were accuracy, f1-score, precision, recall, and ROC_AUC. SHAP explainer was used to explain the predictions made by the XGBoost model. Specifically, SHAP explainer helped in ranking the features based on their importance.

4.4 Results

4.4.1 XGBoost performance

The XGBoost model performance is shown in Table 4.4 using a 10-fold cross validation. Also, a 10-fold cross validation was used to compare the XGBoost performance with other algorithms (e.g., LR, DT, NB, and KNN). As shown in Table 4.4, XGBoost performed almost the best across all metrics (i.e., accuracy, f1-score) among the list of machine learning models.

Since the dataset was imbalanced, we varied the sample size of the trust data as shown in Table 4.5. In the first trial, we used the same sample size for the trust and distrust data. In the second trial, we doubled the sample size of trust data. And in the third trial, we tripled the sample size of trust data. A 10-fold cross validation was used to obtain the results shown in Table 4.5. The XGBoost model had better performance with the increase in the sample size of trust data.

Table 4.4: Performance measure comparison between different models.

Models	Accuracy	ROC_AUC	Precision	Recall	F1-score
Logistic Regression	79.6%	0.68	79.9%	98.9%	88.3%
Decision Tree	75.7%	0.64	82.8%	87.1%	84.9%
Naïve Bayes	75.1%	0.63	80.5%	89.9%	84.9%
KNN	75.6%	0.64	84.3%	84.6%	84.5%
XGBoost	**81.6%**	0.63	83.4%	95.5%	**89.1%**

Table 4.5: Summary of XGBoost classifier performance.

Sample size	Accuracy	ROC_AUC	Precision	Recall	F1-score
Trust (484), Distrust (484)	74.8%	0.75	72.5%	79.9%	75.9%
Trust (968), Distrust (484)	78.5%	0.72	79.3%	91.8%	85.0%
Trust (1452), Distrust (484)	79.8%	0.64	81.2%	94.9%	87.6%
Trust (1745), Distrust (484)	81.6%	0.63	83.4%	95.5%	89.1%

4.4.2 Feature importance

To understand the importance of each feature in predicting trust, SHAP feature importance plot was examined. In addition, we used feature selection to identify the best model performance by adding one feature at a time following the importance ranking of the variables identified in Figure 4.6. We continued this process until the f1-score stopped improving. F1-score was used as our main performance measure since it is a popular evaluation metric for classification analysis. Finally, we found that XGBoost performed the best when a combination of seven features (i.e., mean_HR_max,mean_HRV,number_of_fixations_center,mean_GSR,number_of_fixations_table t, mean_dispersion_tablet, mean_duration_tablet) was used. The SHAP summary plot shown in Fig. 10 has four characteristics including (1) the density represents the distribution of the features in the data, (2) the color shows the range of a particular feature from high (red) to low (blue), (3) the horizontal variation shows the large or small effect of the feature on the prediction, and (4) the vertical ranking represents the importance of the feature. For the feature "number_of_fixations_center", a high number of fixations at the center screen leads to a decrease in trust whereas a low number of fixations at the center screen leads to an increase of trust.

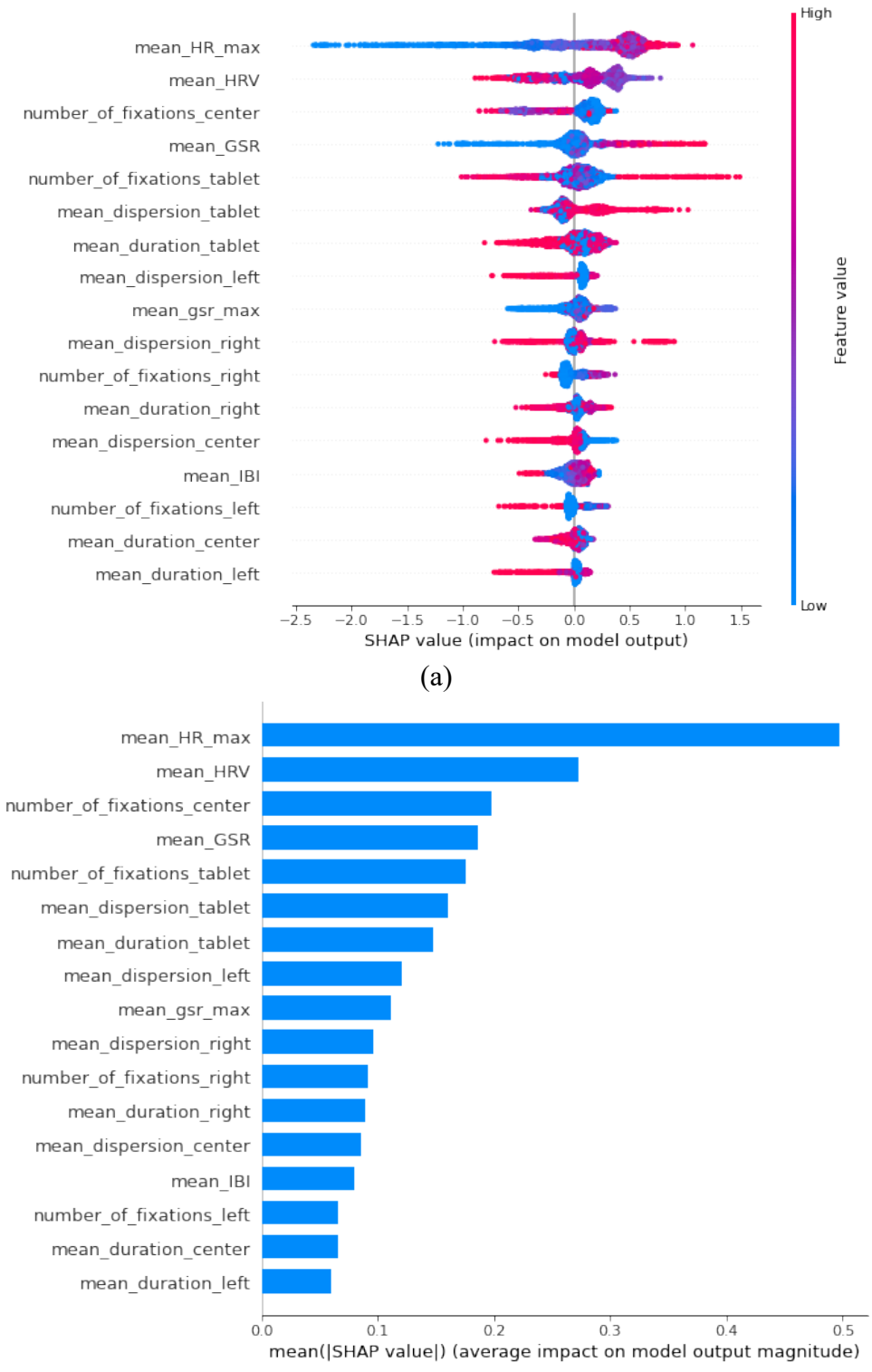


Figure 4.5: (a) SHAP summary plot (b) SHAP feature importance plot.

4.5 Discussion

4.5.1 Contribution and implications

Trust in AVs has been considered as a central factor determining the degree of success in the utilization of AVs. However, there is a gap in the knowledge base that pertains to the practical issue of estimating and calibrating trust in real-time. The main purpose of this study was to develop a prediction model to infer automation trust in real time for different types of drivers to successfully interact with conditional AVs. The contribution of this work is significant because it helped in advancing our understanding of trust as a determining factor to optimize the interaction between the driver and the AV system. As shown in the result section, our proposed framework successfully provides trust predictions in real-time. This study opens the path for more research on the improvement of trust prediction in real-time while using more complex models and exploring more physiological data. In addition, it helps in designing a trust calibration interface by tracking the moments when the driver trust/distrust the AV in real-time using physiological data and machine learning models to control drivers' trust in AVs. Calibration of trust to an appropriate level is considered a design goal to improve the safety and maximize the benefits of AVs while avoiding misuse/disuse of AVs.

4.5.2 Model performance comparison

In this study, we compared the XGBoost classifier model performance with four other machine learning models (e.g., LR, DT, NB, and KNN). As indicated by the results of model accuracy and f1-score, the XGBoost classifier outperformed the other approaches. The results were consistent with previous studies on drivers' trust prediction in AVs (Ayoub et al., 2021).

4.5.3 Effects of features on the trust prediction model

In this study, we proposed a novel method to use a combination of physiological measures such as heart rate activity, galvanic skin responses, and eye-tracking to predict drivers' dynamic trust. Previous studies have used physiological data to measure trust. For instance, Hergeth et al. (2016) showed that eye-tracking can be used to evaluate participants' trust in AVs. A negative relationship was found between monitoring frequency and drivers' self-reported measures which is in line with our obtained results. For instance, for the feature "number_of_fixations_center", a high number of fixations on the center screen decreased participants' trust whereas a low number of fixations at the center screen increased their trust. As for the feature "number_of_fixation_tablet", a high number of fixations on the tablet screen was shown to increase participants' trust more than decreasing their trust. However, it wasn't clear why at a high number of fixations on the tablet screen decreased participants' trust.

For the heart rate related features, a high "mean_HR_max" increased participants' trust while a low "mean_HR_max" decreased their trust. As for the "mean_HRV" feature, a high "mean_HRV" decreased participants' trust while a low "mean_HRV" increased their trust. HRV metric is also a widely used indicator for stress. Assuming that stressful conditions decrease people's trust, our result contradicts what is found in the literature that in stressful conditions the HR was higher and HRV was lower than in control conditions (Held et al. 2021). Thus, to further understand this result a personality analysis, as well as an evaluation of participants' dispositional trust is needed. For instance, if this is a participant first time using a driving simulation, the environment might be stressful for him which increases his HR and decreases his HRV but he can still have high trust in AVs. On the other hand, if a participant had the mentality of distrusting AVs no matter how they behave, his HR can be low and his HRV can be high if he distrusts AVs.

Theoretically, these measures should be applied in real driving situations where low trust indeed induces a high level of stress.

As for the "mean_GSR" feature, a high "mean_GSR" increased participants' trust while a low "mean_GSR" decreased their trust. This aligns with previous studies showing a correlation between GSR and people's trust (Khawaji et al., 2015) (Kumar Akash et al., 2018) (Baig & Kavakli, 2019). Wang et al. (2018) showed that GSR and gaze behavior were negatively associated with self-reported trust.

4.6 Conclusion

In this study, we predicted drivers' trust in real-time using physiological data and a machine learning model. The results showed that the XGBoost classifier model has an accuracy of 81.6% and an F1-score of 89.1% which outperformed other machine learning models. In addition, we identified the most important physiological measures for real-time prediction of trust. Such system can be used in the future to guide the design of an in-vehicle trust calibration warning system to improve people's acceptance and trust in AVs. Future studies should focus on the usage of more complex prediction models such as combining convolution neural network and long short-term memory which has been widely applied in time series modeling due to its effectiveness in modeling temporal relations. In addition, trust varies from person to person as it is influenced by people's personality and dispositional trust levels, therefore it is needed to include personality as a feature in the trust prediction model. Additional improvements to our framework may be conducting the study in more realistic situations that induce more stress like real road testing. Additionally, our methodology needs to be tested in different scenarios with higher NDRT and

scenario complexity. Participants in this study were mostly college students. Future studies should recruit participants from diverse backgrounds, ages, and AV experiences to generalize the analysis.

CHAPTER 5

Calibrating Drivers' Dynamic Situational Trust in Conditional AVs

5.1 Introduction

Undesirable trust levels can dimmish the benefits of using AVs. For instance, one of the leading causes of recent AV crashes (e.g., Tesla's fatal crash in Florida and the Uber AV crash in Arizona) was drivers' overtrust in their AVs' capabilities (Rice, 2019; Kohli & Chadha, 2020). Takeover transitions should be promptly and safely handled when AVs reach their functional limit. Although these crashes probably occurred due to overtrust in the capabilities of conditional AVs (i.e., society of automotive engineers (SAE) levels 2–3), the crashes might reflect a negative first impression with respect to public opinion about AV safety and capabilities. System performance is one of the important factors affecting drivers' dynamic situational trust in AVs (Merritt et al., 2015). In general, a consistently good system performance improves trust and vice versa. However, Merritt et al. (2015) reported the influence of the system performance on participants' overall trust without investigating the trust dynamics. For example, Yin et al. (2019) found that participants' situational trust was affected by the model's stated and observed accuracy, but the researchers did not measure the effect of accuracy on the trust dynamics over time.

One reason to investigate the dynamics of situational trust over time is that people can be potentially 'trapped' in two undesirable trust conditions (i.e., overtrust and undertrust), and it takes time to calibrate their trust with varied system performance. For instance, Schwarz et al. (2019) investigated the effect of varying AV system reliability on participants' trust in conditionally automated driving situations. They found that individuals who did not experience TORs in the initial drive (i.e., in an overtrust condition) did not decrease their trust in AVs in the second drive after experiencing TORs. Hence, it is important to examine the dynamics of situational trust over time. Okamura and Yamada (2020) showed that, by adaptively presenting simple cues, participants in an overtrust precondition were able to calibrate their trust level dynamically over time.

Many researchers have investigated the factors associated with different levels of trust. For instance, Ayoub et al. (2019) identified important factors affecting people's dispositional and learned trust in AVs using an explainable machine learning model. They found that an individual's perceived benefits and risks of AVs, excitement about driving, knowledge of AVs, and eagerness to adopt a new technology were ranked the highest. Zhang et al. (2019) examined the effects of different factors on initial learned trust in AVs using a technology acceptance model and found that perceived safety risk was negatively associated with initial learned trust, while perceived usefulness was positively associated with initial learned trust. They showed that initial learned trust had a high impact on enhancing AV acceptance. In addition, initial learned trust in AVs increased with more interactive automated driving. For example, Gold et al. (2015) found that participants' self-reported (learned) trust was enhanced after experiencing a drive with three TORs. Similarly, Beggiato and Krems (2013) showed that participants' learned trust in adaptive cruise control increased when they were provided with a description of the system capabilities and limitations. However, these studies mostly adopted a snap-shot view that measured trust before

and after the experiment without examining the dynamics of trust over time Guo and Yang (2021), which is critical to calibrating overtrust or undertrust. Hoff & Bashir (2014) identified many factors influencing situational trust, including task difficulty, system performance, perceived risks and benefits, driver workload, experience, attentional capacity, and mood. However, few researchers investigated the dynamics of situational trust in conditional AVs. Hergeth et al. (2016) showed that the participants' self-reported situational trust (SST) dynamically increased from the first to the eighth TOR. Azevedo-Sa et al. (2021) presented a framework for estimating participants' dynamic situational trust due to malfunctions of the AV system. Luo et al. (2020) examined dynamic situational trust using two variables, i.e., level and source of stochasticity. They found that participants' trust decreased significantly due to AV internal errors (e.g., sensor error) versus external (e.g., roadblocks) errors. Okamura and Yamada (2020) examined the effect of the AV system dynamic reliability on participants' situational trust and used trust calibration cues to help improve performance. These studies gained insights into how these variables influenced situational trust.

To understand how situational trust evolves over time, it is important to measure it properly. The majority of existing measures are based on questionnaires, such as the trust scale proposed by Jian et al. (2000), which measures overall trust. One limitation of this scale is that it cannot support the temporal and context-related nature of situational trust. Recently, Holthausen et al. (2020) suggested a short scale for faster implementation based on the trust model suggested by Hoff and Bashir (2015). The suggested scale evaluated situational trust using six items, including trust, performance, NDRT, risk, judgment, and reaction. Trust was also reported to be highly related to individual behaviors (Lee and See 2004). For example, Hergeth et al. (2016)

showed that monitoring frequency during NDRTs in conditional AVs was negatively correlated with three constructs of trust.

To understand how situational trust is built over time, we aimed to investigate in this study the effects of system performance and participant's trust preconditions on the dynamic situational trust during takeover transitions. Both self-reported measures and behavioral measures were considered when determining how situational trust evolves. As a summary, the contributions of this study are:

- We investigated the effects of system performance (i.e., 95%, 80%, and 70%) and trust preconditions (i.e., overtrust and undertrust) on dynamic situational trust in conditionally automated driving simultaneously.
- We used both self-reported and behavioral trust measures to understand the dynamic situational trust in conditionally automated driving.
- The insights obtained from the dynamic situational trust provided important implications on calibrating trust over time in conditionally automated driving.

5.2 Method

5.2.1 Participants

A total number of 42 participants (22 females and 20 males; M = 25.0 years and SD = 5.4 years) located in the United States participated in this study. All the participants were university students, and each had a valid US driver's license. Participants were randomly assigned to one of the two preconditions (i.e., overtrust and undertrust) of the experiment. The average completion time of one session of the study (i.e., 33 minutes) was similar in the two tested preconditions (i.e., overtrust or undertrust).

5.2.2 Apparatus

A survey was designed to evaluate the effects of trust preconditions and system performance on participants' situational trust during takeover scenarios. The survey was developed in Qualtrics (Provo, UT) and administered using Zoom (San Jose, CA). The driving scenarios with takeover requests were created in a virtual environment using Unreal Engine (Cary, NC).

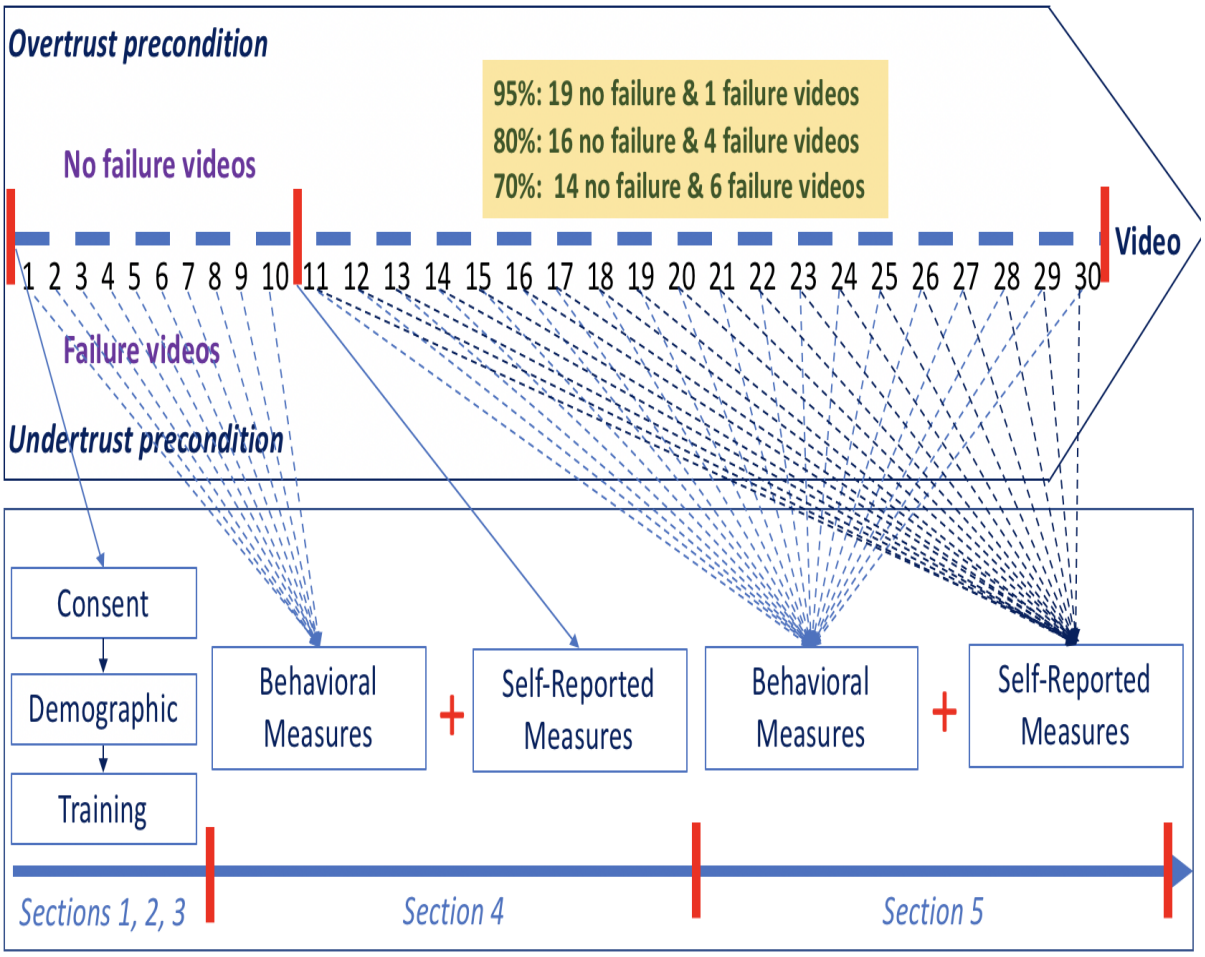


Figure 5.1: Survey procedure.

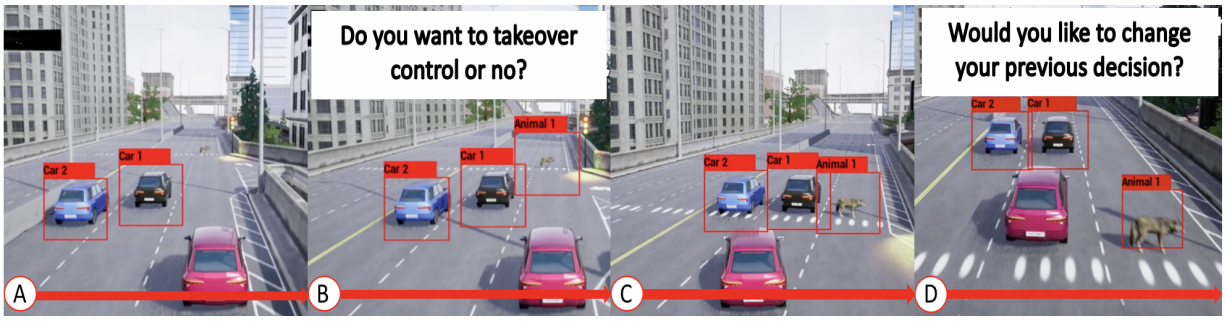


Figure 5.2: Flow of measuring behavioral situational trust during a takeover scenario with no failures.

5.2.3 Experimental design

Independent variables. Our experiment was a 3 (system performance with three accuracy levels: 95%, 80%, and 70%) by 2 (trust preconditions: overtrust and undertrust) mixed-subjects design. The within-subjects variable was the system performance. The between-subjects variable was the trust precondition of the participants by showing them ten takeover scenarios consecutively with successes (to elicit overtrust) or failures (to elicit undertrust).

Dependent variables. We measured participants' dynamic situational trust during the experiment using two types of measures: 1) self-reported measures using the Situational Trust Scale for Automated Driving (STS-AD) (Holthausen et al., 2020) and 2) behavioral measures to capture how often participants agreed/disagree with the AV decision. The STS-AD included six scale items based on the suggested items in the trust model of Hoff and Bashir (Hoff & Bashir, 2014) (i.e., trust, performance, NDRT, risk, judgment, and reaction). The questions asked in this study to evaluate the SST were as follows: 1) I trust the automation in this situation, 2) I would have performed better than the AV in this situation, 3) In this situation, the AV performs well enough for me to engage in other activities, 4) The situation was risky, 5) The AV made an unsafe judgment in this situation, and 6) The AV reacted appropriately to the environment. We measured

the six STS-AD scales with a 7-point Likert scale. This scale was used 21 times during the experiment, where the first measurement was administered after the first 10 videos to manipulate them in a specific precondition (overtrust or undertrust) and the following 20 measurements were conducted right after each of the remaining 20 videos (see Figure 5.1).

As for the behavioral measures, they were proven to be useful in measuring situational trust (e.g., Hergeth et al., 2016). In this work, two questions were used to evaluate participants' behavioral trust adapted from (Yin et al., 2019). The first question "Would you like to takeover control?", was shown before the participants saw how the AV handled the potential takeover transition (see Figure 5.2). Rather than the system-initiated takeover, this scenario imitated the operator-initiated takeover (Wang & Li, 2019) whenever the participant thought it was necessary. The second question "Would you like to change your previous decision?", was asked after the participant saw the vehicle's decision (see Figure 5.2). With this order of questions, we evaluated participants' trust before and after seeing the vehicle behavior. The behavioral trust was measured 30 times in the middle of each video (see Figure 5.1) because we needed multiple measures to calculate agreement fraction and switch fractions in the preconditions. To quantify the behavioral measure of trust, we modified the agreement and switch equations suggested by Yin et al. (2019). The agreement fraction is the number of scenarios for which participants' initial prediction agreed with the vehicle's decision divided by the total number of scenarios presented to the participants (see 5.1). The switch fraction is the number of scenarios for which participants' initial prediction disagreed with their final decision divided by the total number of scenarios (see 5.2).

$$A_k = \frac{\sum_{j=1}^{N}(P_{i_j}^k = D_j^k)}{N}, \tag{5.1}$$

$$S_k = \frac{\sum_{j=1}^{N}(P_{i_j}^k \neq P_{f_j}^k)}{N}, \tag{5.2}$$

where k refers to the kth participant, N is the total number of scenarios, $P_{i_j}^k$ refers to the kth participant's initial prediction in the jth video, $P_{f_j}^k$ refers to the kth participant final decision in the jth video, and D_j^k refers to the vehicle's decision in the jth video for the kth participant. In our study, the participants' initial prediction was based on their willingness to take over control prior to seeing the vehicle's decision. The vehicle's decision was whether the AV would fail or not in that particular scenario. The participants' final decision was based on their willingness to change their previous takeover decision after seeing the vehicle's decision.

5.2.4 Survey design and procedure

The experiment consisted of three sessions (corresponding to three accuracy levels in a counterbalanced order) for one precondition and each session was conducted with a two- day gap. The survey consisted of five sections as illustrated in Figure 5.1. The first section included a consent form. In the second section, the participants filled a set of demographic questions only once during the first session. In the third section, the participants were given a detailed explanation of the study procedure and went through a practice session. The given explanations were about the information provided in the video such as surrounding vehicles and objects and the flow of actions. The explanations helped the participants to think like a passenger to make rational decisions. Then, the participants watched one training video that was different from the testing videos and answered the corresponding questions. In the fourth section, the participants were required to watch 10 videos. If the tested precondition was overtrust, participants were required to watch 10 videos where the AV was able to safely handle the scenarios with no failures. Whereas, if the tested precondition was undertrust, the 10 videos were about AVs failing the driving scenarios. The order of the videos was randomized in the two conditions. In the fifth section, participants watched 20

videos of AV driving scenarios involving potential takeovers (corresponding to the participants initiated a takeover request decision) with a different number of failures, depending on the tested AV accuracy (i.e., 1, 4, and 6 failures in 95%, 80%, and 70% accuracy levels, respectively). For instance, if the tested AV accuracy was 95%, one video showed the AV was able to handle the driving scenario successfully while the remaining 19 videos showed successful driving scenarios with potential takeover transitions. The order of the videos under each tested accuracy was randomized. During the study, each video was paused before showing the decision of the AV, and participants were asked if they would like to take over control. Then, the video was resumed to show the AV's decision. After that, the participants were asked if they would like to change their previous answer. Their decision should be based on the information provided about the surrounding vehicles and the previously watched videos (see Figure 5.2). After each video (starting from the 10th video), they were required to answer the six questions on a 7-point Likert scale in the STS-AD survey.

5.2.5 Scenario design

We investigated drivers' trust in AVs' takeover scenarios by exploring their situational trust in different system performance with different numbers of failures of the vehicle. We conducted a web search to find takeover scenarios based on real experiences. From the web videos, we identified 30 different scenarios that we used to create the failure and non-failure videos. In the failure scenarios, we added bad weather conditions, such as fog and snow to help improve the fidelity. Since adding an adverse weather condition to the scenarios might cue the participants and bias the results, bad weather (e.g., snow weather) was included in three of the nonfailure scenarios. In addition, in the training section, we made clear that AV failure could occur in both bad and good weather conditions depending on the scenarios. A brief description of the scenarios is shown in

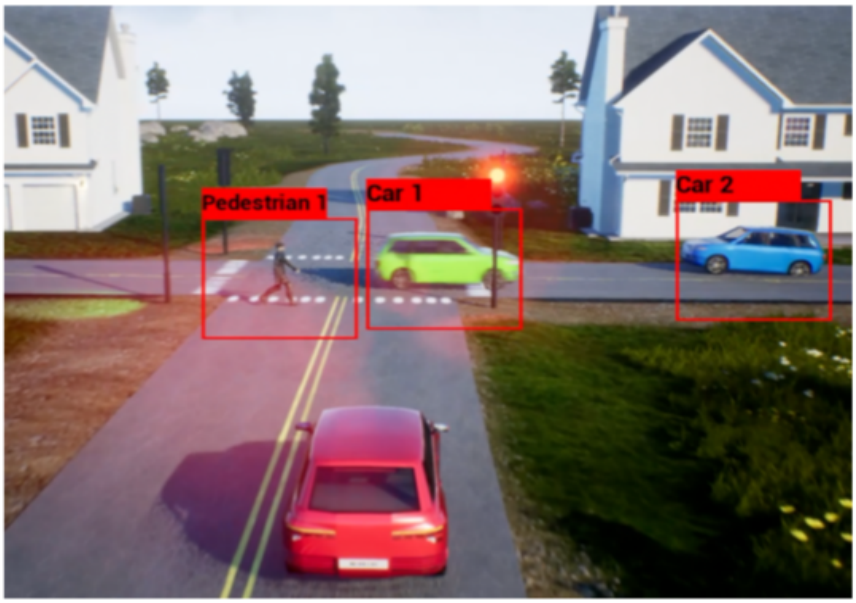
Table 5.1. The failure scenarios were similar to the nonfailure scenarios except that the AV failed to handle the situation. For instance, one of the scenarios showed the AV traveling on a highway with a broken down truck parked at the side of the road. In the nonfailure scenario, the AV safely detected the truck and changed lanes before hitting the truck [see Fig. 5.3(a)]. However, in the failure scenario, the AV failed to detect the truck due to foggy conditions and crashed into the truck [see Fig. 5.3(b)].

Table 5.1: Description of the created scenarios.

	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**
Location	Highway or Street	Highway or Street	Highway or crossroad	Crossroad	Highway
Target	Pedestrian or Animal or Construction	Truck or Vehicle	Vehicle	Red light	Vehicle
Action	Road crossing	Broken down truck or vehicle	Stopping at a stop sign	Stopping at a red light	Merging lanes

5.2.6 Data analysis

Statistical analysis was conducted using SPSS statistics software (IBM, New York City, NY, USA). An analysis of variance (ANOVA) was used to analyze the effects of system performance and trust preconditions on dynamic situational trust. The alpha level was set at 0.05 for all the statistical tests. The ANOVA assumptions, including normality and homogeneity of variance, were not violated for either overtrust or undertrust preconditions. Pairwise comparisons were performed with Bonferroni correction.



(a)

(b)

Figure 5.3: Takeover scenario with (a) no failure and (b) failure.

5.3 Results

5.3.1 Manipulation check

We compared the self-reported situational trust (SST) in the precondition stage (i.e., the manipulation stage with 10 videos of consecutive failures or successes) with that before the manipulation and that after the manipulation. Before the manipulation stage, a question (i.e., in general, how much do you trust an autonomous vehicle) was asked in the demographic section of the survey.

In the overtrust precondition, we found that participants' SST (M = 5.327, $S.E.$= .113, see the 1st video in Figure 12b) was significantly higher (F (1, 20) = 12.910, p = .002) than the overall trust level before manipulation (M = 4.286, $S.E.$= .269) and was significantly higher than that in the testing stage by aggregating the last 20 videos in the 95% (p = .001), 80% (p = .000), and 70% (p = .000) accuracy levels. Furthermore, in the undertrust precondition, participants' self-reported situational trust (SST) did not have significant differences (F (2, 40) = 1.709, p = .194) among the three accuracy conditions. Similarly, for the overtrust precondition (see Figure 12b), participants' SST had no significant differences (F (2, 40) = .342, p = .712) among the three accuracy levels. These results showed that participants were calibrated to the overtrust and undertrust preconditions.

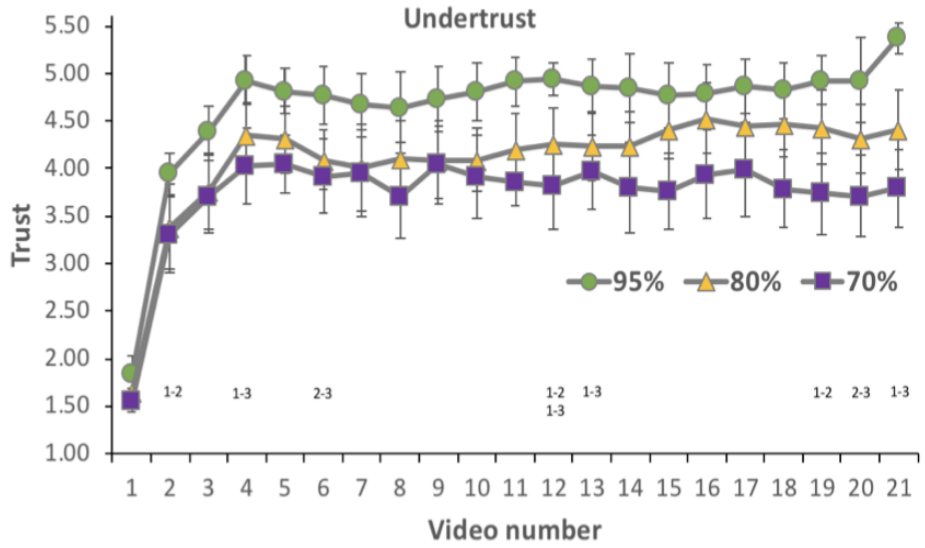
However, we were not able to measure behavioral situational trust (BST) in the demographic section of the survey. In the manipulated precondition (i.e., in the first 10 videos), for the agreement fraction, using a two-way mixed ANOVA, we did not find any significant differences among the three accuracy levels (F (2, 80) = 1.429, p = .246), two preconditions (F (1, 40) = .041, p = .218), or interaction between accuracy and preconditions (F (2, 80) = 1.554, p = .842). We further compared the agreement fraction between the first 10 videos and the last 20

videos, there were no significant differences found. For the switch fraction, using a two-way mixed ANOVA, we did not find any significant differences among three accuracy levels (F (2, 80) = 2.669, p = .076) or interaction between accuracy and preconditions (F (2, 80) = 1.836, p = .167). However, we found that those in the undertrust precondition had a significantly higher switch fraction than those in the overtrust precondition (F (1, 40) = 9.484, p = .004). This might be due to the fact that each participant was asked to participate in the experiment for three sessions in the same precondition. We further compared the switch fraction between the first 10 videos and the last 20 videos, we did not find any significant differences.

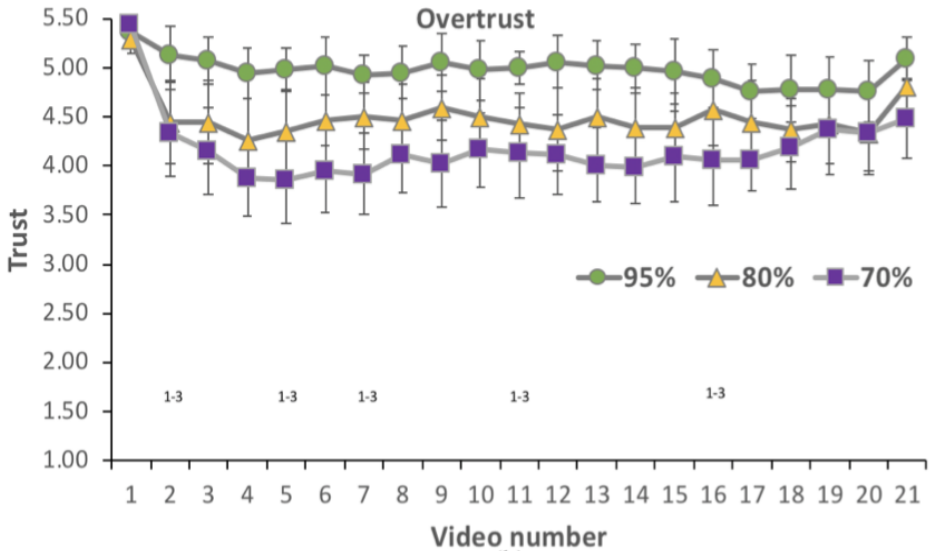
5.3.2 Self-reported situational trust

Due to a relatively small number of participants, we calculated the mean of the SST over all the 21 participants from the second video (in Figure 5.4, i.e., the 11th video in the experiment) with a smoothing function by using a window length of 5 videos. Note that the first video in the figure represents the average results of the first 10 videos that the participants watched to manipulate them in an undertrust or overtrust condition, which was not smoothed with the other following videos. Figure 5.5 illustrates the mean SST for all 21 participants with three tested accuracy levels under the two trust preconditions (i.e., overtrust and undertrust).

A mixed two-way ANOVA showed that the main effect of accuracy was significant (F (2, 80) = 144.794, p = .000) whereas the main effect of trust preconditions was not significant (F (1, 40) = .915, p = .344) (see Figure 5.5). There was no significant interaction effect between the trust preconditions and the tested accuracy levels (F (2, 80) = 0.269, p = .765). The pairwise comparison showed a significant difference between 95% and 80% accuracies (p = .000), 95% and 70%



(a)



(b)

Figure 5.4: Overall mean and standard error of self-reported situational trust (SST) measured by the STS-AD six scales for all the participants at different accuracy levels and trust preconditions. (a) Undertrust precondition. (b) Overtrust precondition. Along the x-axis the accuracy levels having significant differences of pairwise comparisons are indicated with number pairs. “1” indicates 95%, “2” indicates 80%, and “3” indicates 70%.

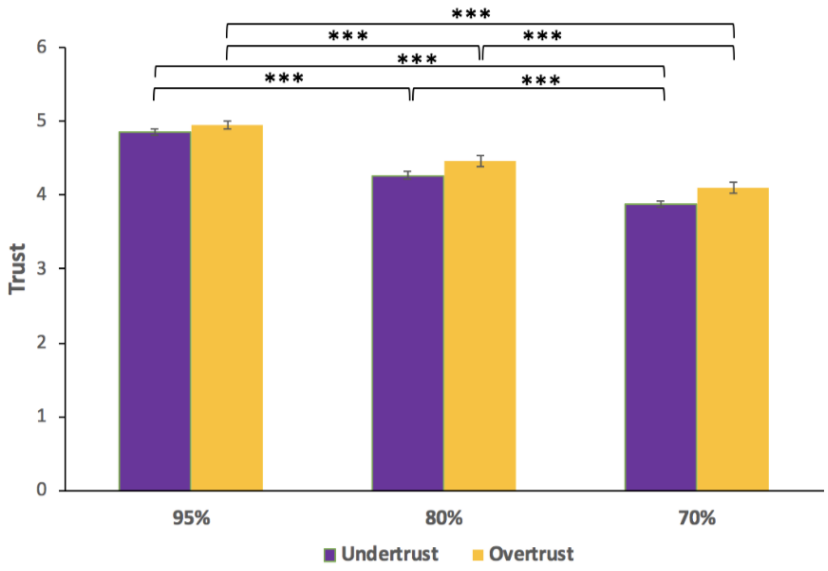


Figure 5.5: Mean self-reported situational trust at different accuracy and trust precondition levels with standard errors, where"*" indicates $p < 0.05$, "**" indicates $p < 0.01$, and "***" indicates $p < 0.001$.

accuracies (p = .000), and 80% and 70% accuracies (p = .000) (see Figure 5.5). Due to the significant main effect of accuracy levels, we also conducted a pairwise comparison among three accuracy levels at each video as illustrated along the x-axis in Figure 12 for both undertrust and overtrust preconditions. Significant differences were labeled by the numbers in the figure, where "1" indicates 95%, "2" indicates 80%, and "3" indicates 70%.

Furthermore, we analyzed each of the STS-AD trust scale separately using a two-way ANOVA as shown in Figure 5.6 and Figure 5.7. We found significant main effects of accuracy levels for all the six questions of the STS-AD scale (all p = .000). In addition, we conducted a pairwise comparison at each video as illustrated along the x-axis in Figure 5.6 and Figure 5.7. Whenever a significant difference existed at each video, they were labeled by the numbers in the figure. There was a significant main effect for trust precondition for Q1 (F (1, 40) = 4.293, p =

.045) and pairwise comparison showed significant differences when comparing 95% vs. 95% (p = .000), 80% vs. 80% (p = .007), and 70% vs. 70% (p = .000) between two preconditions.

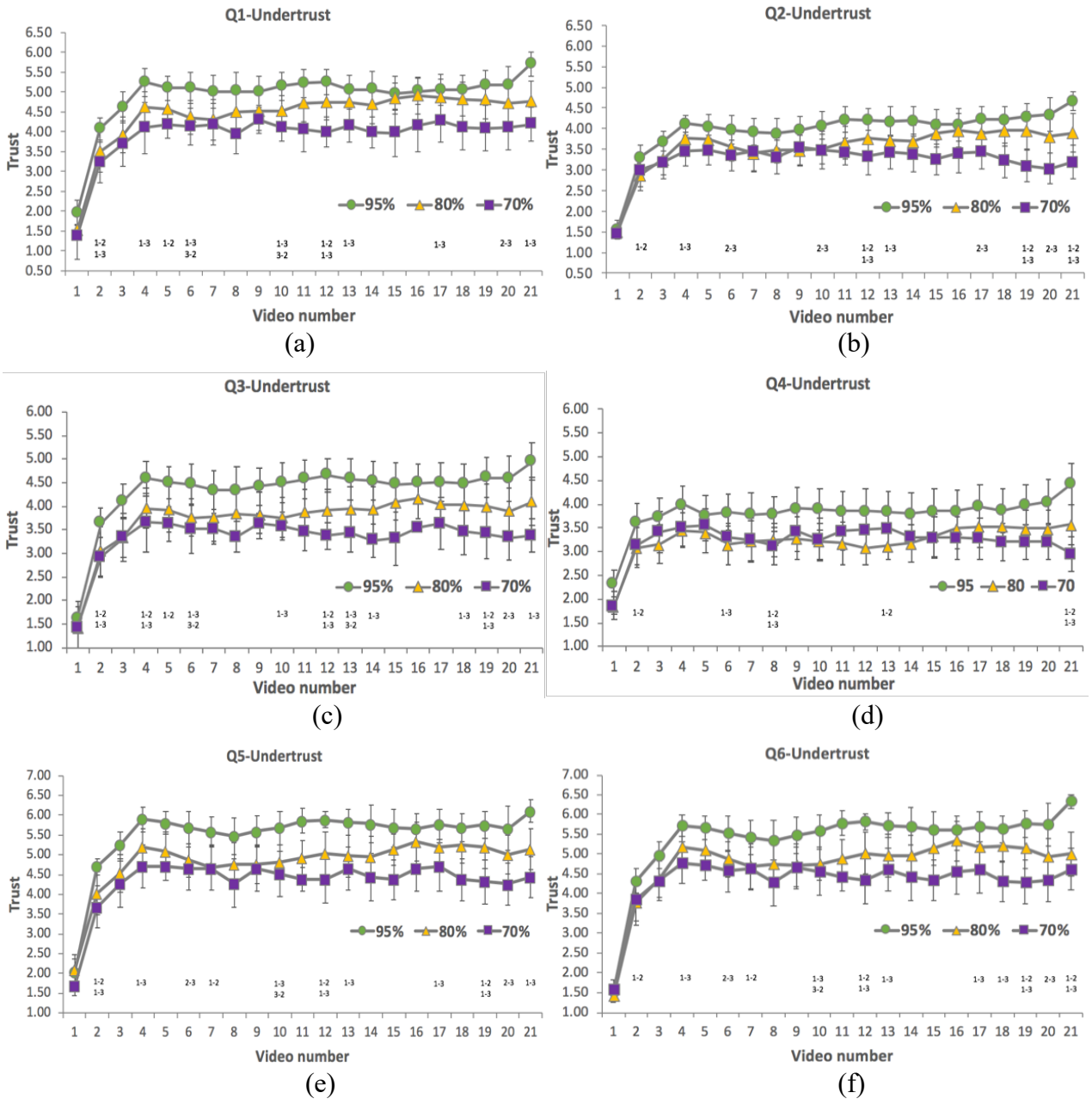


Figure 5.6: Mean measures of the STS-AD six scales for all the participants in the undertrust precondition and at different accuracy levels. (a) Q1. (b) Q2. (c) Q3. (d) Q4. (e) Q5. (f) Q6. Along the x-axis, the accuracy levels having a significant difference in pairwise comparisons are indicated with number pairs. “1” indicates 95%, “2” indicates 80%, and “3” indicates 70%.

As for Q2, Q3, Q4, and Q6 there was no significant difference between two trust conditions. As for Q5, there was a significant main effect for trust precondition (F (1, 40) = 4.530, p = .040) and significant differences were found when comparing 80% vs. 80% (p = .019) and 70% vs. 70% (p = .009) between the two preconditions. We grouped the results according to the transition condition of the videos. We have identified four patterns of failure occurrences including one failure (i.e., 113 occurrences in the undertrust precondition and 123 occurrences in the overtrust precondition), two consecutive failures (i.e., 31 occurrences in the undertrust precondition and 27 occurrences in the overtrust precondition), three consecutive failures (i.e., 5 occurrences in 25 the undertrust precondition and 8 occurrences in the overtrust precondition), and four consecutive failures (i.e., 2 occurrences in the undertrust precondition and 2 occurrences in the overtrust precondition). The situations where the failure scenario occurred as the first or last video in the 20 videos sequence were removed from this analysis. The case with four consecutive failures was not analyzed since it only occurred 32 two times in the undertrust and overtrust preconditions. In the case of one failure video, there was a significant difference in the SST level between the failure video and the previous non-failure (i.e., undertrust F (1, 112) = 817.466, p = .000; overtrust F (1, 122) = 754.016, p = .000)) and between the failure video and the following non-failure video (i.e., undertrust F (1, 112) = 968.681, p = .000; overtrust F (1, 122) = 662.665, p = .000) (see Figure 5.8a). In the case of two consecutive failure videos, there was a significant difference in the SST level between the average of the two consecutive failure videos and the previous non-failure video (i.e., undertrust F (1, 30) = 253.144, p = .000; overtrust F (1, 26) = 339.807, p = .000) and between the average of the two consecutive failure videos and the following non- failure video (i.e., undertrust F (1, 30) = 267.374, p = .000; overtrust F (1, 26) = 405.087, p = .000) (see Figure 5.8b). In the case of three consecutive failure videos, there was a significant

difference in the SST level between the average of the three consecutive failure videos and the previous non- failure video (i.e., undertrust F (1, 4) = 19.033, p = .012; overtrust F (1, 7) = 49.860, p = .000) and between the average of the three consecutive failure videos and the following non-failure video in the overtrust precondition (i.e., F (1, 14) = 113.370, p = .000). There was a marginal significant difference between the average of the three consecutive failure videos and the following non-failure video in the undertrust precondition (i.e., F (1, 4) = 6.459, p = .064) (see Figure 5.8c).

5.3.3 Behavioral situational trust

Agreement and switch fractions were used in this work as BST measures of the participants' dynamic situational trust. As shown in Figure 5.10a, for the agreement fraction, the main effect of accuracy was not significant on BST (F (2, 80) = .497, p = .61) whereas the main effect of precondition was significant (F (1, 40) = 8.553, p = .006). There was no significant interaction effect between the trust precondition and the tested accuracy levels (F (2, 80) = .483, p = .619). Pairwise comparison showed that the agreement fraction was significantly higher in the overtrust condition than that in the undertrust condition when comparing 95% vs. 95% (p = .029) and 70% vs. 70% (p = .006) and marginally higher in the overtrust condition than that in the undertrust condition when comparing 80% vs. 80% (p = .087). For the switch fraction as shown in Figure 5.10b, the main effect of accuracy levels was not significant on BST (F (2, 80) = 1.888, p = .158) whereas the main effect of precondition was significant (F (1, 40) = 10.053, p = .003). There was no significant interaction effect between the trust precondition and the tested accuracy levels (F (2, 80) = 0.657, p = .521). Pairwise comparison showed that the switch fraction in the undertrust condition was significantly higher than in the overtrust condition when comparing 95%

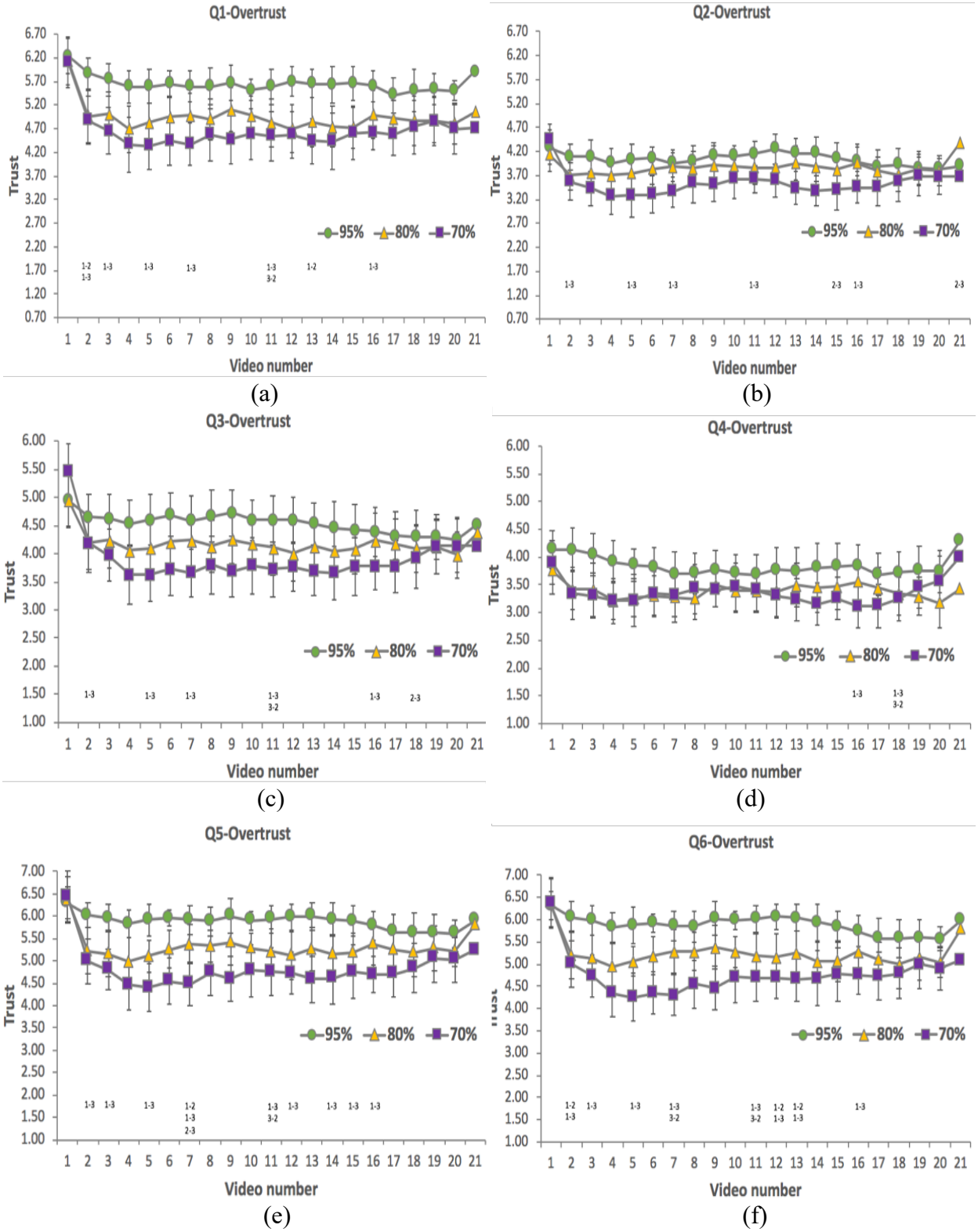
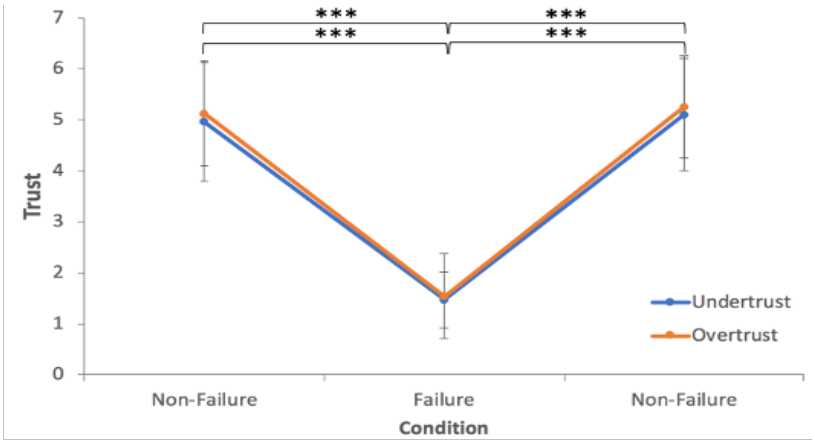
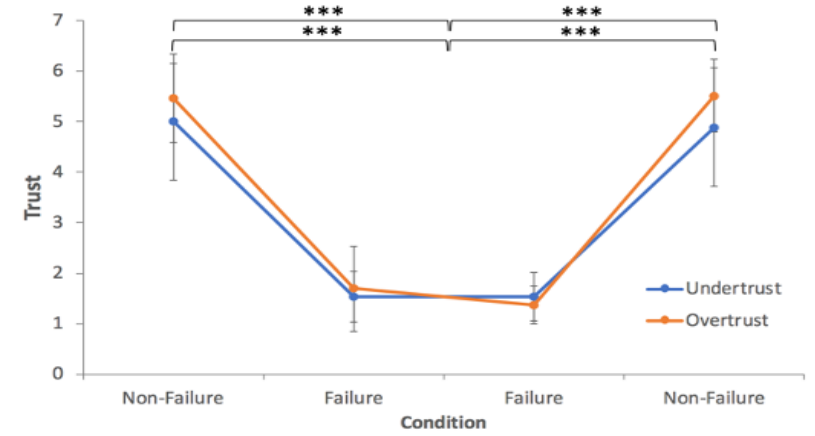


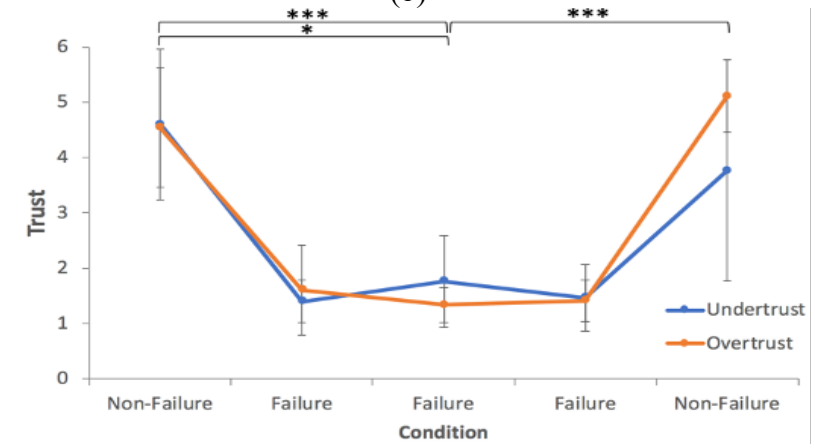
Figure 5.7: Mean measures of the STS-AD six scales for all participants in the overtrust precondition and at different accuracy levels. (a) Q1. (b) Q2. (c) Q3. (d) Q4. (e) Q5. (f) Q6. Along the x-axis, the accuracy levels having a significant difference in pairwise comparisons are indicated with number pairs. “1” indicates 95%, “2” indicates 80%, and “3” indicates 70%.



(a)



(b)



(c)

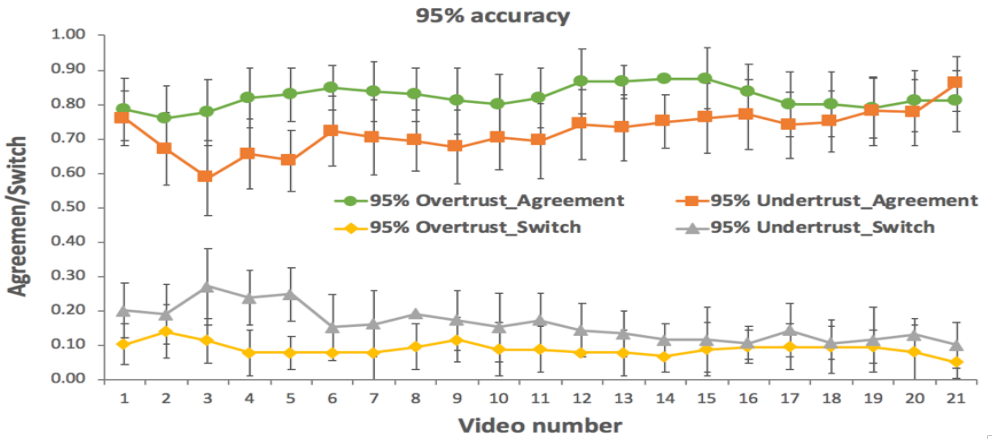
Figure 5.8: Mean measure of the STS-AD six scales in the overtrust and undertrust precondition at different consecutive failures occurrences with standard deviation, where '*' indicates $p < 0.05$,'**' indicates $p < 0.01$, and '***' indicates $p < 0.001$. (a) 1 failure. (b) 2 failures. (c) 3 failures.

vs. 95% (p = .023) and 80% vs. 80% (p = .004), and on the borderline of significance when comparing 70% vs. 70% (p = .050) between the two preconditions. Due to the significant main effect of preconditions, we also conducted a pairwise comparison at each video as illustrated in Figure 5.9 for both undertrust and overtrust preconditions for the agreement and switch fractions. Significant differences were labeled by '*'.

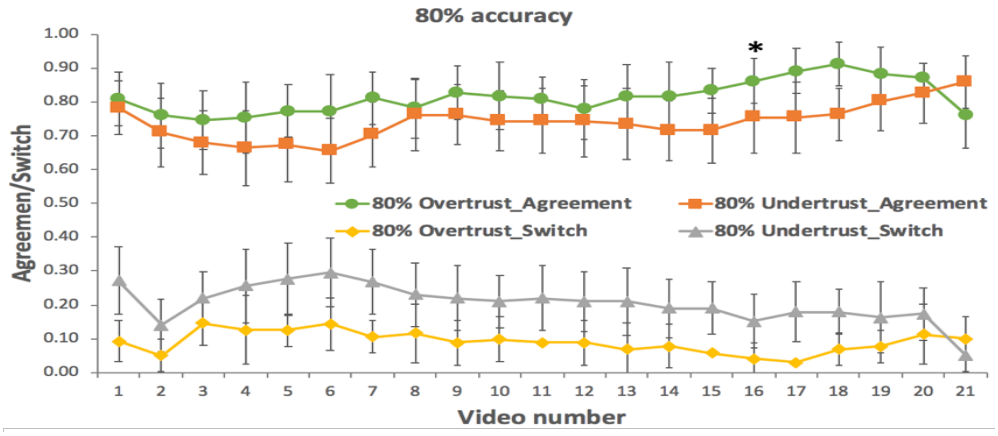
5.4 Discussion

5.4.1 Self-reported situational trust

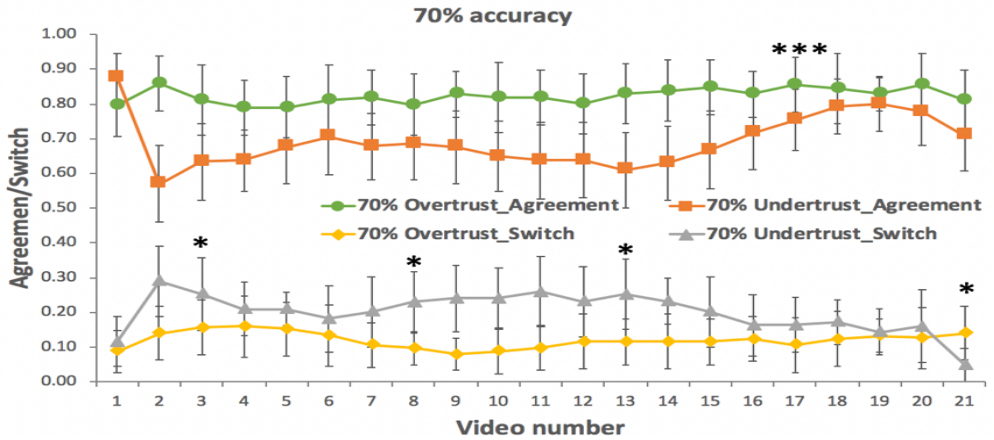
In this study, we aimed to understand the effects of trust preconditions and system performance on dynamic situational trust in conditional AVs. First, by letting the participants watch 10 videos of takeover scenarios with or without failures, we were able to manipulate them in an overtrust or undertrust condition by examining the SST measure. However, we were not able to measure BST before the experiment so that we were not able to test BST for manipulation check for a pre-test comparison. We were not able to find any significant differences between the preconditions and test conditions. This indicated the BST was not good at measuring trust levels manipulated by vehicle performance levels, which was also evidenced by the insignificant main effect of accuracy levels. For the SST measure, it was quickly calibrated to the different accuracy levels from their corresponding trust pre-conditions (see Figure 5.4). We noticed that participants' average SST level between the 10th and 11th video increased from around 1.563 to 3.300 for 70% (p = .000), from 1.643 to 3.383 for 80% (p = .000), and from around 1.841 to 3.947 for 95% (p = .000) accuracy levels (see Figure 5.4a). Whereas in Figure 4b, we noticed that participants'



(a)

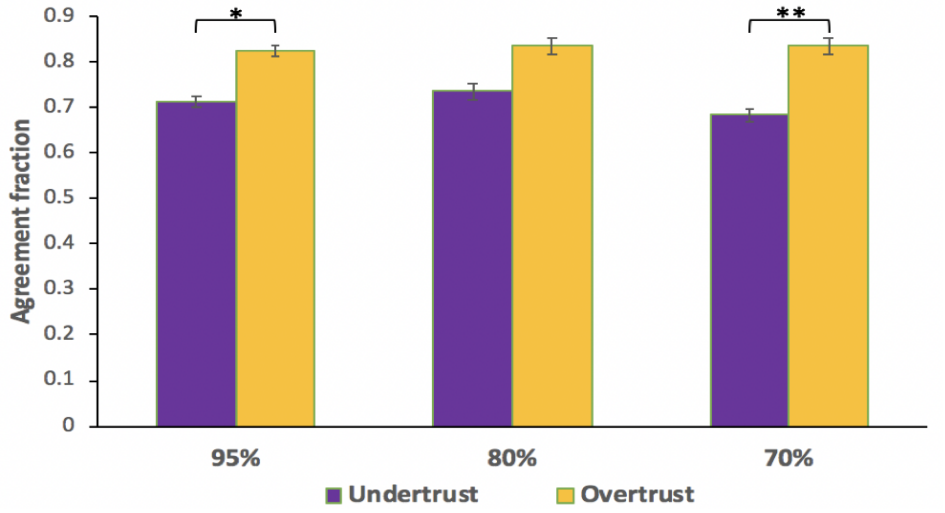


(b)

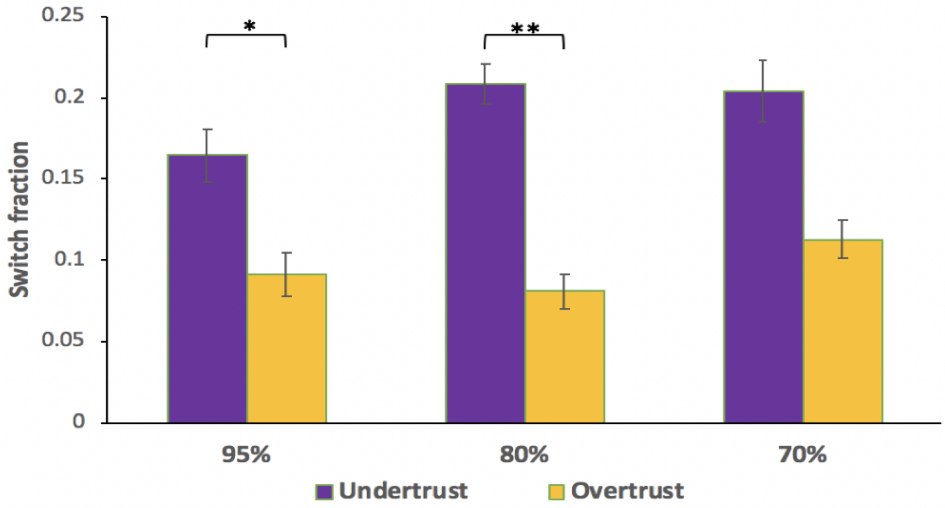


(c)

Figure 5.9: Overall mean and standard error of behavioral situational trust measured by the agreement and switch fractions for all the participants at different accuracy levels and trust preconditions. (a) 95% accuracy. (b) 85% accuracy. (c) 70% accuracy. Note that the first video in the graph represents the average results of the first 10 videos that the participants watched in order to manipulate them in an undertrust or overtrust condition, where '*' indicates $p < 0.05$,'**' indicates $p < 0.01$, and '***' indicates $p < 0.001$.



(a)



(b)

Figure 5.10: (a) Mean agreement fraction with standard error and (b) Mean switch fraction with standard error at different accuracy levels and in the overtrust and undertrust preconditions, where "*" indicates $p < 0.05$, "**" indicates $p < 0.01$, and "***" indicates $p < 0.001$.

average SST level at the 10th video decreased significantly from 5.294 to 4.450 for 80% ($p = .010$) and from 5.452 to 4.333 for 70% ($p = .001$) accuracy levels. As for 95% accuracy, there was no significant difference ($p = .341$) in trust between the 10th and 11th video. It showed that the participants were able to quickly calibrate their SST based on the performance of the AV. This effect seemed to be more prominent in the undertrust precondition than the overtrust precondition by comparing Figure 5.4a and Figure 5.4b for the first several SST scores. In the undertrust

precondition, participants took more time (i.e., until the 4th video after the precondition) to calibrate their trust after watching the 10 failure videos. However, in the overtrust precondition, the participant calibrated their trust faster (i.e., until around the 2nd video after the precondition). This seemed to be consistent with previous findings (Parasuraman and Manzey 2010) that failures led to a deep drop in trust and the recovery was slow. We noticed an increase in trust for the last video in the undertrust and overtrust preconditions at the three tested accuracies (see Figure 5.4). Due to the random order assigned in Qualtrics, the majority of the scenarios for the last video were non-failure scenarios.

Second, our findings showed that the participants were able to perceive the system performance at different accuracy levels and to adjust their SST based on the system performance (see Figure 5.6 and Figure 5.7), even though the accuracy of the system was not presented to the participants. We noticed that no matter what the trust precondition of the participants was, they always had a higher SST in the 95% compared to 80% and 70%. This is consistent with previous studies (e.g., Hergeth et al., 2016 and Beggiato and Krems 2013) that participants had the capabilities of learning the performance of the AV dynamically and calibrating their SST level over time by understanding the capabilities and limitations of the AV over time. This also indicated that the SST levels closely reflected the actual performance of the AVs, which could potentially help avoid misuse and abuse of conditional AVs.

5.4.2 Behavioral situational trust

Agreement and switch fractions were used in this study as BST measures of the participants' dynamic situational trust. First, in Figure 5.10a, we noticed that the agreement fraction in the overtrust precondition was significantly higher than in the undertrust precondition for all the tested accuracy levels. Whereas in Figure 5.10b, the switch fraction was significantly

higher in the undertrust condition than that in the overtrust precondition. However, there was no significant main effect for different accuracy levels. Hence, the SST and BST were complementary to each other in that the SST measure was sensitive to system performance while the BST was sensitive to trust preconditions. The agreement fraction indicated participants' reliance on the AV to a large extent (automation reliance) while the switch fraction indicated participants' rejection of the AV's decisions (self-reliance) (Dzindolet et al., 2003). When the participants went through the first 10 videos with all successes, they tended to overtrust the vehicle was at least as reliable as manual operations if not more reliable than manual operations. This made them more likely to rely on the automation and agree with the system's decisions. On the contrary, when the participants were in the undertrust condition, they tended to have a low consistency with the system's decisions, which made them switch their initial prediction about what the system would 42 handle the driving scenarios. Such results were supported by previous findings that a high level of trust was associated with more reliance on automation and vice versa (Wickens et al. 2015). The difficulty in reducing such automation bias (e.g., automation reliance and self-reliance) might also explain why different levels of system performance did not play a role in BST when they were in an overtrust or undertrust preconditions.

5.4.3 Comparison between SST and BST

The obtained results showed that the SST and BST were complementary to each other. However, Yin et al. (2019) showed a similar trust pattern in the analysis of self-reported and behavioral measures. Their self-reported and behavioral measures were different from the ones used in our work. Murtin et al. (2018) showed that both self-reported and behavioral measures are correlated with the expected trustworthiness, but behavioral trust additionally captures the willingness to cooperate during a specific interaction. They concluded that these two measures are

related but they should be considered as complementary. Our results showed that the SST measure was only sensitive to system performance since the main effect of accuracy was significant. By analyzing the items of the STS-AD scale, we noticed that only Q1 (i.e., I trust the automation in this situation) and Q5 (i.e., The AV made unsafe judgment in this situation) had a significant main effect of trust precondition. This result might be caused by the low fidelity of the system that made it hard to estimate the performance, NDRT, risk, and reaction to the environment. Another reason might be that the participants took a short period of time to develop similar SST levels between the overtrust and undertrust preconditions. Furthermore, our results showed that the BST was only sensitive to trust preconditions since the main effect of precondition was significant. In our experiment, the SST and BST were not measured at the same time which could also be the reason for not obtaining the same trust pattern between these two measures. BST was measured during the scenario while SST was measured after the scenario was done.

5.4.4 Implications

Trust plays an important role in adopting and proper use of AVs and different constructs of trust could have different nature. Situational trust is dynamically evolving depending on multiple factors in the human-AV interaction process. Our study showed that the participants were able to calibrate their SST to the real performance of the AV over time. This indicates that clearly showing the capabilities and limitations of the AV can help drivers to quickly calibrate their trust level (Lee and See 2004). However, the calibrated trust level was not influenced by their trust preconditions. However, the effect of trust preconditions was reflected by the BST measures, which could be considered as learned trust in AVs in this study. If such learned trust is inconsistent with the system performance, automation bias can occur. Thus, it is important to consider drivers' learned trust in AVs when designing calibration systems as these preconditions could potentially

influence their calibrated situational trust over time. We also showed two types of situational trust measures, i.e., SST and BST, complemented each other and the inconsistency between them calls for further investigation in the reliability of different measures.

5.5 Conclusion and future work

In this work, we investigated the effects of system performance and participants' trust preconditions on the dynamic situational trust in conditional AVs. The dynamic situational trust was measured using self-reported and behavioral measures and the participants were able to adjust their SST levels dynamically to be consistent with the performance of the AV. However, such results were moderated by their trust preconditions measured by BST levels. Such insights revealed important implications for designing a calibration system for conditional AVs. Our study also has limitations, which can be left for future research. First, this study was conducted in a low-fidelity experiment setup with a small sample size. Future studies 64 should be conducted in a high-fidelity driving simulator or even in a naturalistic driving environment with a larger and diverse sample size to see if there will be consistent results. Second, trust was mainly evaluated in takeover scenarios in conditional AVs using SST with the STS-AD scales and BST with the agreement and switch fractions. Thus, further analyses are needed to explore trust in other types of scenarios with other possible measures, such as eye-tracking data (Hergeth et al., 2016). Third, failure scenarios happened only in bad weather conditions, which might bias participants' trust evaluation.

Also, for the first 10 videos, the average SST was calculated only once at the end of the 10th video to save time. In the undertrust precondition, after watching 10 failure videos, participants might doubt that the AV was not capable. Their knowledge level regarding the AV capabilities could play a role in their trust formation, which was not studied in this paper. Fourth,

this study mainly investigated the effects of trust preconditions and system performance on dynamic situational trust. Future studies could potentially include other factors, such as cognitive workload, and explore the two other layers of trust (e.g., learned trust and dispositional trust) to gain a complete understanding of the effects of system performance and trust precondition.

CHAPTER 6

Conclusion

6.1 Summary of research achievements

To address the difficulties related to trust estimation in autonomous vehicles and fill in the current research gaps, the objectives of this dissertation were summarized as follows:

1) Develop a computational model to predict drivers' dispositional and initial learned trust using machine learning and survey data.
2) Develop a computational model to predict situational trust using physiological measurements in real time in conditional AVs.
3) Investigate the effect of system performance and people's trust preconditions on the dynamic situational trust during takeover to provide implications for designing an alerting system to calibrate people's trust in conditional AVs.

To meet objective 1, we attempted to predict trust in AVs with both accuracy and explainability using XGBoost and SHAP models. To predict trust in AVs, we conducted an online survey to collect various variables that were related to participants' trust in AVs. The survey data were then used to train and test the XGBoost model. To help better understand the XGBoost model,

SHAP was used to explain the trust predictions by identifying the most important predictor variables, by examining their interaction effects, and by illustrating individual explanation cases. Compared with previous trust predictions studies, our proposed method combines the benefits of XGBoost and SHAP with good explainability and predictability of the trust model.

To meet objective 2, we predicted drivers' trust in real time using physiological data and machine learning models. The results showed that the XGBoost classifier model has an accuracy of 81.6% and an F1-score of 89.1% which outperformed other machine learning models. In addition, we identified the most important physiological measures for real time prediction of trust. Such a system can be used in the future to guide the design of an in-vehicle trust calibration warning system to improve people's acceptance and trust in AVs.

To meet objective 3, we investigated the effects of system performance and participants' trust preconditions on the dynamic situational trust in conditional AVs. The dynamic situational trust was measured using self-reported and behavioral measures and the participants were able to adjust their self-reported situational trust levels dynamically to be consistent with the performance of the AV. However, such results were moderated by their trust preconditions measured by behavioral situational trust levels. Such insights revealed important implications for designing a calibration system for conditional AVs.

6.2 Intellectual merit and broad impact

The proposed research will expand the knowledge about trust in AVs and thus it will contribute to improving people's acceptance of AVs. Findings from chapter 3 helped in identifying the most important factors affecting trust in AVs. And compared with previous trust prediction

models, our proposed method improved the predictability and explainability of modeling trust. Next, in chapter 4 we collected physiological data to assess situational trust using machine learning models. Our target was to predict drivers' trust in real time with high accuracy. Finally, in chapter 5 we investigated the effects of trust precondition and system performance on the dynamic situational trust. Also, we showed how self-reported and behavioral measures can be used to measure dynamic situational trust. Findings from this research provided important implications for designing a trust calibration system for AVs. The society in general will benefit from our obtained findings to increase safety, acceptance, and usage of AVs.

Intellectual Merit: The findings from this research will help foster users' trust in AVs and they can also be used by AV manufacturers to design a trust calibration system in AVs whenever undertrust or overtrust is detected in the system by incorporating machine learning models with human factors methods in AVs.

Broader Impact: Being able to predict drivers' trust by understanding what factors drive people's trust in AV will improve their acceptance of AVs. Specifically, predicting drivers' dynamic trust can help in detecting inappropriate calibration of trust and thus contribute to increased safety while being in an AV.

6.3 Future work

Although this dissertation helped in enhancing our understanding of trust in autonomous Vehicles, there are several limitations that should be considered in the future.

First**,** future study should focus on the usage of more complex prediction models such as combining convolution neural networks and long short-term memory which has been widely applied in time series modeling due to its effectiveness in modeling temporal relations.

Second, trust varies from person to person as it is influenced by people's personality and dispositional trust levels, therefore it is needed to include personality as a feature in the trust prediction model.

Third, to see the robustness of the obtained results, the conducted studies should be replicated in more realistic situations that induce more stress like naturalistic driving.

Fourth, our methodologies need to be tested in different scenarios with higher mental workload and scenario complexity.

Finally, participants in the conducted studies were mostly college students. Future studies should recruit participants from diverse backgrounds, age, and AV experiences to generalize the analysis to see if there will be consistent results.

APPENDICES

APPENDIX A

Questionnaire for Experiment 2 in Chapter 4

A.1 Trust evaluation self-report question

Please indicate your trust level after this encounter:

Extremely Low Trust					Neither Low/High Trust					Extremely High Trust
0	1	2	3	4	5	6	7	8	9	10

Trust level

APPENDIX B

Questionnaires for Experiment 3 in Chapter 5

B.1 Situational trust scale for automated driving

Please answer the following questions based on the watched video:

	Strongly disagree 1	2	3	Neither agree nor disagree 4	5	6	Strongly agree 7
I trust the automation in this situation	○	○	○	○	○	○	○
I would have performed better than the AV in this situation	○	○	○	○	○	○	○
In this situation, the AV performs well enough for me to engage in other activities	○	○	○	○	○	○	○
The situation was risky	○	○	○	○	○	○	○
The AV made unsafe judgement in this situation	○	○	○	○	○	○	○
The AV reacted appropriately to the environment	○	○	○	○	○	○	○

REFERENCES

Akash, K., Wan-Lin Hu, Reid, T., & Jain, N. (2017). Dynamic modeling of trust in human-machine interactions. *2017 American Control Conference (ACC)*, 1542–1548. https://doi.org/10.23919/ACC.2017.7963172

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1–20. https://doi.org/10.1145/3132743

Ayoub, J., Mason, B., Morse, K., Kirchner, A., Tumanyan, N., & Zhou, F. (2020). Otto: An Autonomous School Bus System for Parents and Children. *CHI Extended Abstracts*. https://doi.org/10.1145/3334480.3382926

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Modeling dispositional and initial learned trust in automated vehicles with predictability and explainability. *Transportation Research Part F: Traffic Psychology and Behaviour*, *77*, 102–116. https://doi.org/10.1016/j.trf.2020.12.015

Ayoub, J., & Zhou, F. (2020). Investigating drivers' trust in autonomous vehicles' decisions of lane changing events. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1274–1278. https://doi.org/10.1177/1071181320641303

Ayoub, J., Zhou, F., Bao, S., & Yang, X. J. (2019). From manual driving to automated driving: A review of 10 years of autoui. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 70–90.

Azevedo-Sa, H., Jayaraman, S.K., Esterwood, C. T., X. Yang, X. J., Robert, L. P., & Tilbury, D. M. 2021. Real-Time Estimation of Drivers' Trust in Automated Driving Systems. *International Journal of Social Robotics*, 1-17. https://doi.org/10.1007/s12369-020-00694-1.

Baig, M. Z., & Kavakli, M. (2019). A Survey on Psycho-Physiological Analysis & Measurement Methods in Multimodal Systems. *Multimodal Technologies and Interaction*, *3*(2), 37. https://doi.org/10.3390/mti3020037

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Bearth, A., & Siegrist, M. (2016). Are risk or benefit perceptions more important for public acceptance of innovative food technologies: A meta-analysis. *Trends in Food Science & Technology*, *49*, 14–23. https://doi.org/10.1016/j.tifs.2016.01.003

Beggiato, Matthias, & Josef F. Krems. 2013. The Evolution of Mental Model, Trust and Acceptance of Adaptive Cruise Control in Relation to Initial Information. *Transportation Research. Part F, Traffic Psychology and Behaviour* 18 (May): 47–57.

Benedek, M., & Kaernbach, C. 2010. A Continuous Measure of Phasic Electrodermal Activity. *Journal of Neuroscience Methods* 190 (1): 80–91.

Berkovsky, S., Taib, R., Koprinska, I., Wang, E., Zeng, Y., Li, J., & Kleitman, S. (2019). Detecting Personality Traits Using Eye-Tracking Data. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300451

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cho, J.-H., Chan, K., & Adali, S. (2015). A Survey on Trust Modeling. *ACM Computing Surveys*, *48*, 1–40. https://doi.org/10.1145/2815595

Choi, J., & Ji, Y. G. (2015). Investigating the Importance of Trust on Adopting an Autonomous Vehicle. *International Journal of Human-Computer Interaction*, *31*, 150709133142005. https://doi.org/10.1080/10447318.2015.1070549

Davenport, R. B., & Bustamante, E. A. (2010). Effects of False-Alarm vs. Miss-Prone Automation and Likelihood Alarm Technology on Trust, Reliance, and Compliance in a Miss-Prone Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(19), 1513–1517. https://doi.org/10.1177/154193121005401933

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 251–258. https://doi.org/10.1109/HRI.2013.6483596

Doney, P. M., Cannon, J. P., & Mullen, M. R. (1998). Understanding the Influence of National Culture on the Development of Trust. *The Academy of Management Review*, *23*(3), 601–620. JSTOR. https://doi.org/10.2307/259297

Dong, X., Victor, U., & Qian, L. (2020). Two-path Deep Semi-supervised Learning for Timely Fake News Detection. *ArXiv:2002.00763 [Cs]*. http://arxiv.org/abs/2002.00763

Du, N., Zhou, F., Pulver, E. M., Tilbury, D. M., Robert, L. P., Pradhan, A. K., & Yang, X. J. (2020). Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation Research Part C: Emerging Technologies*, *112*, 78–87. https://doi.org/10.1016/j.trc.2020.01.006

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Edmonds, E. (2019, March 14). Three in Four Americans Remain Afraid of Fully Self-Driving Vehicles. AAA NewsRoom. https://newsroom.aaa.com/2019/03/americans-fear-self-driving-cars-survey/

Ekman, F., Johansson, M., & Sochor, J. (2018). Creating Appropriate Trust in Automated Vehicle Systems: A Framework for HMI Design. *IEEE Transactions on Human-Machine Systems*, *48*(1), 95–101. https://doi.org/10.1109/THMS.2017.2776209

Elrod, L. (2014). National Highway Traffic Safety Administration. In *Encyclopedia of Transportation: Social Science and Policy* (Vol. 1–4, pp. 965–966). SAGE Publications, Inc. https://doi.org/10.4135/9781483346526

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, *77*, 167–181. https://doi.org/10.1016/j.tra.2015.04.003

Gempler, K. S. (1997). Display of Predictor Reliability on a Cockpit Display Traffic Information. [Master's thesis, University of Illinois at Urbana-Champaign]. https://apps.dtic.mil/dtic/tr/fulltext/u2/a366236.pdf

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in Automation – Before and After the Experience of Take-over Scenarios in a Highly Automated Vehicle. *Procedia Manufacturing*, *3*, 3025–3032. https://doi.org/10.1016/j.promfg.2015.07.847

Guo, Y., & Yang, X. J. 2021. Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics* 13 (8): 1899–1909

Guo, B., Ding, Y., Yao, L., Liang, Y., & Yu, Z. (2019). The Future of Misinformation Detection: New Perspectives and Trends. *ArXiv:1909.03654 [Cs]*. http://arxiv.org/abs/1909.03654

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences*, *116*(16), 7684–7691. https://doi.org/10.1073/pnas.1805770115

Held, J., Andreea V., Christine W., Nadine M., & Christoph F.. 2021. Heart Rate Variability Change during a Stressful Cognitive Task in Individuals with Anxiety and Control Participants. *BMC Psychology* 9 (1): 44

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep Your Scanners Peeled: Gaze Behavior as a Measure of Automation Trust During Highly Automated Driving. *Human Factors*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744

Hoff, K. A., & Bashir, M. (2014). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*. https://doi.org/10.1177/0018720814547570

Hogervorst, M. A., Brouwer, A.-M., & van Erp, J. B. F. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00322

Holmes, J. G. (1991). Trust and the appraisal process in close relationships. In *Advances in personal relationships: A research annual, Vol. 2.* (pp. 57–104). Jessica Kingsley Publishers

Holthausen, B. E., Wintersberger, P., Walker, B. N., & Riener, A. (2020). Situational Trust Scale for Automated Driving (STS-AD): Development and Initial Validation. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 40–47). Association for Computing Machinery. https://doi.org/10.1145/3409120.3410637

Hoogendoorn, M., Jaffry, S. W., & Treur, J. (2008). Modeling Dynamics of Relative Trust of Competitive Information Agents. In M. Klusch, M. Pěchouček, & A. Polleres (Eds.), *Cooperative Information Agents XII* (pp. 55–70). Springer. https://doi.org/10.1007/978-3-540-85834-8_7

Jacobs, S. C., Friedman, R., Parker, J. D., Tofler, G. H., Jimenez, A. H., Muller, J. E., Benson, H., & Stone, P. H. (1994). Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research. *American Heart Journal*, *128*(6, Part 1), 1170–1177. https://doi.org/10.1016/0002-8703(94)90748-X

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The Measurement of the Propensity to Trust Automation. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, Augmented and Mixed Reality. Applications and Case Studies* (pp. 476–489). Springer International Publishing. https://doi.org/10.1007/978-3-030-21565-1_32

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Johnson, T. (2012, July 25). *Enhancing Safety Through Automation*. http://onlinepubs.trb.org/onlinepubs/conferences/2012/Automation/presentations/Johnson.pdf

Jonker, C. M., & Treur, J. (1999). Formal Analysis of Models for the Dynamics of Trust Based on Experiences. In F. J. Garijo & M. Boman (Eds.), *Multi-Agent System Engineering* (pp. 221–231). Springer. https://doi.org/10.1007/3-540-48437-X_18

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies*, *96*, 290–303. https://doi.org/10.1016/j.trc.2018.07.001

Khawaji, A., Zhou, J., Chen, F., & Marcus, N. (2015). Using Galvanic Skin Response (GSR) to Measure Trust and Cognitive Load in the Text-Chat Environment. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 1989–1994. https://doi.org/10.1145/2702613.2732766

Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, *44*(2), 544–564. https://doi.org/10.1016/j.dss.2007.07.001

Kohli, P., & Chadha, A. (2020). Enabling Pedestrian Safety using Computer Vision Techniques: A Case Study of the 2018 Uber Inc. Self-driving Car Crash. *ArXiv:1805.11815 [Cs]*, *69*, 261–279. https://doi.org/10.1007/978-3-030-12388-8_19

Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. *Bagnara S., Tartaglia R., Albolino S., Alexander T., Fujita Y. (Eds) Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). IEA 2018*, *vol 823*. https://doi.org/10.1007/978-3-319-96074-6_2

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50-80. http://doi.org/10.1518/hfes.46.1.50_30392

Lee, J., Kim, N., Imm, C., Kim, B., Yi, K., & Kim, J. (2016). A Question of Trust: An Ethnographic Study of Automated Cars on Real Roads. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 201–208. https://doi.org/10.1145/3003715.3005405

Lewicki, R., & Brinsfield, C. (2011). Framing trust: Trust as a heuristic. *Framing Matters: Perspectives on Negotiation Research and Practice in Communication*, 110–135.

Lewis, J. D., & Weigert, A. (1985). Trust as a Social Reality. *Social Forces*, *63*(4), Pages 967-985. https://doi.org/10.1093/sf/63.4.967

Li, M., Holthausen, B. E., Stuck, R. E., & Walker, B. N. (2019). No Risk No Trust: Investigating Perceived Risk in Highly Automated Driving. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '19*, 177–185. https://doi.org/10.1145/3342197.3344525

Liu, X., Tredan, G., & Datta, A. (2011). A Generic Trust Framework For Large-Scale Open Systems Using Machine Learning. *Computing Research Repository - CORR*, *30*. https://doi.org/10.1111/coin.12022

López, J., & Maag, S. (2015). Towards a Generic Trust Management Framework Using a Machine-Learning-Based Trust Model. *2015 IEEE Trustcom/BigDataSE/ISPA*, *1*, 1343–1348. https://doi.org/10.1109/Trustcom.2015.528

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777

Luo, R., Chu, J., & Yang, X. J. 2020. Trust Dynamics in Human-AV (Automated Vehicle) Interaction. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7. CHI EA '20. New York, NY, USA: Association for Computing Machinery

Madhani, K., Khasawneh, M. T., Kaewkuekool, S., Gramopadhye, A.K., & Melloy, B. J. 2002. Measurement of Human Trust in a Hybrid Inspection for Varying Error Patterns. *Proceedings of the Human Factors and Ergonomics Society. Annual Meeting Human Factors and Ergonomics Society. Meeting* 46 (3): 418–22

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

McCrae, R. R., & Costa, P. T. (2003). *Personality in Adulthood: A Five-factor Theory Perspective*. Guilford Press

Menon, N., Pinjari, A., Zhang, Y., & Zou, L. (2016, January 1). *Consumer Perception and Intended Adoption of Autonomous Vehicle Technology – Findings from a University Population Survey*

Merritt, S. M., Huber, K., LaChapell-Unnerstall, J., & Lee, D. (2014). *Continuous Calibration of Trust in Automated Systems*. MISSOURI UNIV-ST LOUIS. https://apps.dtic.mil/sti/citations/ADA606748

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors*, *57*(1), 34–47

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral Measurement of Trust in Automation: The Trust Fall. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*, 1849–1853. https://doi.org/10.1177/1541931213601422

MUIR, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. https://doi.org/10.1080/00140139408964957

Murtin, F., Fleischer, L., Siegerink, V., Aassve, A., Algan, Y., Boarini, R., … & Smith, C. 2018. Trust and Its Determinants. OECD Statistics Working Papers. Organisation for Economic Co-Operation and Development (OECD). https://doi.org/10.1787/869ef2ec-en

NHTSA2010. (2010). Traffic Safety Facts 2010 A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. National Highway Traffic Safety Administration National Center for Statistics and Analysis U.S. Department of Transportation. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811659

Numan, J. H. (1998). *Knowledge-based systems as companions: Trust, human computer interaction and complex systems*. Undefined. https://www.semanticscholar.org/paper/Knowledge-based-systems-as-companions%3A-Trust%2C-human-Numan/afb2b16ea898a8fd5ec603a38e69c1d742e75e35

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLOS ONE*, *15*(2), e0229132. https://doi.org/10.1371/journal.pone.0229132

Paden, B., Cap, M., Yong, S. Z., Yershov, D., & Frazzoli, E. (2016). A Survey of Motion Planning and Control Techniques for Self-driving Urban Vehicles. *ArXiv:1604.07446 [Cs]*. http://arxiv.org/abs/1604.07446

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 9

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 51–55

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Pavlou, P. A. (2003). Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model (SSRN Scholarly Paper ID 2742286). Social Science Research Network. https://papers.ssrn.com/abstract=2742286

Peters, E., Västfjäll, D., Gärling, T., & Slovic, P. (2006). Affect and decision making: A "hot" topic. *Journal of Behavioral Decision Making*, *19*(2), 79–85. https://doi.org/10.1002/bdm.528

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual Differences in the Calibration of Trust in Automation. *Human Factors*, *57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Radlmayr, J., Gold, C., Lorenz, L., Farid, M., & Bengler, K. (2014). How Traffic Situations and Non-Driving Related Tasks Affect the Take-Over Quality in Highly Automated Driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 2063–2067. https://doi.org/10.1177/1541931214581434

Rajaonah, B., Tricot, N., Anceaux, F., & Millot, P. (2008). The role of intervening variables in driver–ACC cooperation. *International Journal of Human-Computer Studies*, *66*(3), 185–197. https://doi.org/10.1016/j.ijhcs.2007.09.002

Raue, M., D'Ambrosio, L. A., Ward, C., Lee, C., Jacquillat, C., & Coughlin, J. F. (2019). The Influence of Feelings While Driving Regular Cars on the Perception and Acceptance of Self-Driving Cars: Feelings and Self-Driving Cars. *Risk Analysis*, *39*(2), 358–374. https://doi.org/10.1111/risa.13267

Rice, D. (2019). The Driverless Car and the Legal System: Hopes and Fears as the Courts, Regulatory Agencies, Waymo, Tesla, and Uber Deal with this Exciting and Terrifying New Technology. *Journal of Strategic Innovation and Sustainability*, *14*(1), 134–146

Rice, S. (2009). Examining Single- and Multiple-Process Theories of Trust in Automation. *The Journal of General Psychology*, *136*(3), 303–322. https://doi.org/10.3200/GENP.136.3.303-322

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 101–108. https://doi.org/10.1109/HRI.2016.7451740

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*(1), 76–87. https://doi.org/10.1518/001872007779598082

Rudin-Brown, C. M., & Parker, H. A. (2004). Behavioural adaptation to adaptive cruise control (ACC): Implications for preventive strategies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *7*(2), 59–76. https://doi.org/10.1016/j.trf.2004.02.001

Ruijten, P. A. M., Terken, J. M. B., & Chandramouli, S. N. (2018). Enhancing Trust in Autonomous Vehicles through Intelligent User Interfaces That Mimic Human Behavior. *Multimodal Technologies and Interaction*, *2*(4), 62. https://doi.org/10.3390/mti2040062

Sanchez, J. (2006). *Factors that affect trust and reliance on an automated aid*. https://smartech.gatech.edu/handle/1853/10485

Schaefer, A., Nils, F., Sanchez, X., & Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion*, *24*(7), 1153–1172. https://doi.org/10.1080/02699930903274322

Schoettle, B., & Sivak, M. (2016). Motorists' Preferences for Different Levels of Vehicle Automation*: 2016* (SWT-2016-8). Article SWT-2016-8. https://trid.trb.org/view/1480408

Seppelt, B. D. (2009). Supporting operator reliance on automation through continuous feedback [Doctor of Philosophy, University of Iowa]. https://doi.org/10.17077/etd.90t9uqnh

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, *1*. https://doi.org/10.1038/s41562-017-0202-6

Subramanian, R., Wache, J., Abadi, M. K., Vieriu, R. L., Winkler, S., & Sebe, N. (2018). ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing*, *9*(2), 147–160. https://doi.org/10.1109/TAFFC.2016.2625250

Walker, F., Wang, J., Martens, M. H., & Verwey, W. B. (2019). Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 401–412. https://doi.org/10.1016/j.trf.2019.05.021

Walker, G. H., Stanton, N. A., & Salmon, P. (2016). Trust in vehicle technology. *International Journal of Vehicle Design*, *70*(2), 157. https://doi.org/10.1504/IJVD.2016.074419

Wang, M., Hussein, A., Rojas, R. F., Shafi, K., & Abbass, H. A. (2018). EEG-Based Neural Correlates of Trust in Human-Autonomy Interaction. *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 350–357. https://doi.org/10.1109/SSCI.2018.8628649

Wang, S., & Li, Z. (2019). Exploring causes and effects of automated vehicle disengagement using statistical modeling and classification tree based on field test data. *Accident Analysis & Prevention*, *129*, 44–54

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*(4), 352–367. https://doi.org/10.1080/14639220110110306

Wickens, Christopher D., Justin G. Hollands, Simon Banbury, & Raja Parasuraman. 2015. Engineering Psychology and Human Performance. https://doi.org/10.4324/9781315665177

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12. https://doi.org/10.1145/3290605.3300509

Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., & Zhang, W. 2019. The Roles of Initial Trust and Perceived Risk in Public's Acceptance of Automated Vehicles. *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2018.11.018

Zhou, F., Lei, B., Liu, Y., & Jiao, R. J. (2017). Affective parameter shaping in user experience prospect evaluation based on hierarchical Bayesian estimation. *Expert Systems with Applications*, *78*, 1–15. https://doi.org/10.1016/j.eswa.2017.02.003

Zhou, F., Qu, X., Helander, M. G., & Jiao, J. (Roger). (2011). Affect prediction from physiological measures via visual stimuli. *International Journal of Human-Computer Studies*, *69*(12), 801–819. https://doi.org/10.1016/j.ijhcs.2011.07.005

Zmud, J., N.Sener, I., & Wagner, J. (2016). *Consumer Acceptance and Travel Behavior Impacts of Automated Vehicles*. https://static.tti.tamu.edu/tti.tamu.edu/documents/PRC-15-49-F.pdf