

1 **Modeling Dispositional and Initial learned Trust in**
2 **Automated Vehicles with Predictability and**
3 **Explainability**

4 ¹Jackie Ayoub, ²X. Jessie Yang, ¹Feng Zhou

5 ¹Department of Industrial and Manufacturing Systems Engineering, University of Michigan-
6 Dearborn, Dearborn, MI, USA

7 ²Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor,
8 Ann Arbor, MI, USA

9 Accepted to be published in *Transportation Research Part F: Traffic Psychology and Behaviour*,
10 Dec. 23, 2020

11 Corresponding author:

12 Feng Zhou, 4901 Evergreen Road, Dearborn, MI 48128, Email: fezhou@umich.edu

13 **ABSTRACT**

14 Technological advances in the automotive industry are bringing automated driving closer
15 to road use. However, one of the most important factors affecting public acceptance of
16 automated vehicles (AVs) is the public’s trust in AVs. Many factors can influence
17 people’s trust, including perception of risks and benefits, feelings, and knowledge of
18 AVs. This study aims to use these factors to predict people’s dispositional and initial
19 learned trust in AVs using a survey study conducted with 1175 participants. For each
20 participant, 23 features were extracted from the survey questions to capture his/her
21 knowledge, perception, experience, behavioral assessment, and feelings about AVs.
22 These features were then used as input to train an eXtreme Gradient Boosting (XGBoost)
23 model to predict trust in AVs. With the help of SHapley Additive exPlanations (SHAP),
24 we were able to interpret the trust predictions of XGBoost to further improve the
25 explainability of the XGBoost model. Compared to traditional regression models and
26 black-box machine learning models, our findings show that this approach was powerful
27 in providing a high level of explainability and predictability of trust in AVs,
28 simultaneously.

1 *Keywords:* Trust prediction, XGBoost, SHAP explainer, Feature importance, Automated
2 vehicles

3 **1. Introduction**

4 Automated vehicles (AVs) are the next technological revolution due to their advantages
5 (e.g., safer, cleaner, and more efficient) compared to manual driving (Elrod, 2014).

6 Although the automated driving systems will one day handle the whole driving task and
7 allow drivers to do non-driving related tasks (SAE, 2018), the public seems reluctant to
8 adopt the technology. A survey study showed that only 37% of their participants would
9 probably buy an AV (J. D. Power, 2012). Menon (2015) showed that 61.5% of
10 Americans were not willing to use AVs. Such results may be partially due to Uber's and
11 Tesla's crashes involving automated driving, which have shaken consumers' trust in
12 AVs. A recent AAA study (Edmonds, 2019) revealed that three out of four Americans
13 were afraid of using AVs. Therefore, Shariff et al. (2017) and Bansal et al. (2016)
14 concluded that the key barriers to the adoption of AVs are psychological rather than
15 technological, and the most critical one is probably trust.

16 According to Lee and See (2004), trust is defined as “*the attitude that an agent will help*
17 *achieve an individual's goals in a situation characterized by uncertainty and*
18 *vulnerability*”. Researchers identified many factors affecting people's trust in AVs.

19 Ayoub et al. (2019) summarized the factors affecting trust into three categories, including
20 1) human-related factors (i.e., culture, age, gender, experience, workload, and knowledge
21 about AVs), 2) automation-related factors (i.e., reliability, uncertainty, and user
22 interface), and 3) environmental-related factors (i.e., risk, reputation of original
23 equipment manufacturers). Hoff and Bashir (2015) identified three layers of variability
24 in human-automation trust, including dispositional trust, situational trust, and learned
25 trust. Dispositional trust indicates people's enduring tendency to trust, situational trust
26 measures the construct related to trust dynamics during human-automation interaction in
27 specific contexts, while learned trust is related to how past experiences in automated
28 systems influence individuals' current level of trust in automation.

29 Estimating trust in AVs is challenging, especially when the majority of the public does
30 not have much interaction experience with AVs. Raue et al. (2019) suggested that
31 people's experience in manual driving should potentially shape their trust assessment in

1 AVs. Along the same line, Abe et al. (2017) made use of manual driving characteristics
2 (e.g., speeds and time headway) to investigate driver's trust in automated driving in terms
3 of overtaking and passing patterns. Researchers have used linear and logistic regression
4 models (Raue et al., 2019) and other machine learning (e.g., Support Vector Machines
5 (SVMs)) methods to investigate the effects of various factors on trust in AVs (López and
6 Maag, 2015; Liu et al., 2011). While regression models are limited to predict people's
7 trust in AVs, machine learning models, such as SVMs, are capable of providing better
8 predictability at the price of their explainability, i.e., they are considered as black boxes
9 (Rudin, 2019; Adadi and Berrada, 2018).

10 To fill the research gaps, this study proposes a method that provides both good
11 predictability and explainability of trust in AVs, using eXtreme Gradient Boosting
12 (XGBoost) (Chen and Guestrin, 2016) and SHapley Additive exPlanations (SHAP)
13 (Lundberg et al., 2020). By trust in AVs, we mainly measure dispositional and initial
14 learned trust prior interacting with an AV as in (Hoff and Bashir, 2015), because we
15 collected the data through a survey without providing chances for the participants to
16 interact with AVs. First, XGBoost is an ensemble machine learning model based on a
17 large number of decision trees that use an optimized gradient boosting system. It has the
18 advantage to perform parallel processing, to approximate greedy search, and to improve
19 the learning process in the smallest amount of time without overfitting. It was
20 demonstrated that XGBoost had the best prediction performance and processing time
21 compared to gradient boosting and random forest (Chen and Guestrin, 2016). Second, in
22 order to improve the explainability of the XGBoost model, we used SHAP (Lundberg et
23 al., 2020), which uses a game theoretic approach that explains the output of a machine
24 learning model. It combines optimal credit allocation with local and global explanations
25 using the classic Shapley values from game theory and their related extensions. In
26 summary, this study made the following contributions: (1) We proposed a machine
27 learning model that estimates people's dispositional and initial learned trust in AVs with
28 good predictability and explainability; (2) We identified critical factors affecting people's
29 dispositional and initial learned trust in AVs; (3) We identified the main effects and the
30 interaction effects between the critical factors that explain trust prediction.

1 **2. Related Work**

2 **2.1. Factors Affecting Trust in AVs**

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

7 **Perception of Risks:** Risk is considered to be an intrinsic aspect affecting trust, i.e.,
8 when the perceived risk of a situation is high, a higher level of trust is needed to rely on
9 AV's decisions (Numan, 1998; Kim et al., 2008; Pavlou, 2003). Therefore, it is essential
10 to consider factors associated with risks in AVs when evaluating trust (Rajaonah et al.,
11 2008). Zmud et al. (2016) reported that safety risks due to system failures were the major
12 concerns of using AVs. Moreover, Menon et al. (2016) showed that one third of US
13 drivers were worried about the risks of misusing their private AV data. Li et al. (2019)
14 demonstrated that the perceived risks and trust in an AV were affected by introductory
15 information related to system reliability. Therefore, it is important to include risk
16 perception and an appropriate level of information regarding AVs to evaluate trust in the
17 early stages of driver-vehicle interactions.

18 **Perception of Benefits:** Many researchers have found that the perception of benefits is
19 related to improving trust in AVs, which subsequently leads to user acceptance and
20 adoption (Choi and Ji, 2015; Bearth and Siegrist, 2016). One of the major benefits
21 associated with AVs is to reduce vehicle crashes and to save lives. Vehicle crashes lead
22 to injury of 2.2 million Americans each year (NHTSA, 2010) and the cost associated with
23 these crashes is around \$300 billion (Bearth and Siegrist, 2016). Therefore, the safety
24 enhancement behind AVs should be focused on creating crash-less vehicles (Johnson,
25 2012; Fagnant and Kockelman, 2015; Paden et al., 2016). As a matter of fact, human
26 factors were reported to be the cause of 90% of crashes and the death of over 30 thousand
27 Americans per year (Elrod, 2014). AVs are accurate and quicker to react in case of
28 emergency since they can optimize the decision before taking any actions. Aside from

1 improving safety, AVs can bring other social benefits, including reducing congestions,
2 fuel consumption, and CO2 emission (Fagnant and Kockelman, 2015), and so on.

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

18 **Effect of Feelings:** Trust is composed of two components: a cognitive component and an
19 affective component (Lewicki and Brinsfield, 2011; Cho et al., 2015). The cognitive
20 component is based on judgements, beliefs, competence, stability, and expectations while
21 the affective component is based on positive and negative emotions that shape our trust
22 (Lewis and Weigert, 1985). For example, positive emotions were found to improve
23 takeover performance in AVs, which further led to trust in AVs (Du et al., 2020) while
24 negative emotions, such as concerns and worries, made parents trust automated school
25 buses less (Ayoub et al., 2020). Furthermore, Peters et al. (2006) explained that affect
26 influenced our stored knowledge, which further guided our acceptance and trust. Hence,
27 emotion can be used to evaluate trust. According to Hancock and Nourbakhsh (2019), the
28 majority of drivers had no chance to experience AVs yet. Thus, this inexperience makes
29 it harder to evaluate their trust in the system. Raue et al. (2019) suggested that feelings
30 related to people's experience in driving could shape their perception of risks, benefits,

1 and trust in AVs. Specifically, Baumeister et al. (2001) showed that negative emotions
2 were more significant in shaping judgment than positive ones.

3 **2.2. Modeling Techniques of Trust in AVs and Automation**

4 Many researchers used questionnaires (Körber, 2018) and behavioral methods (Miller et
5 al., 2016; Jessup et al., 2019) to evaluate trust in automation and in AVs. For instance,
6 Körber, (2018) built a multidimensional model to measure trust in automation using a
7 survey study. The model was composed of 19 parameters, including reliability,
8 understandability, propensity to trust, familiarity, and intentions. Miller and Perkins
9 (2010) developed a survey to study trust in automation by focusing on 5 components of
10 trust including competence, predictability, dependability, consistency, and confidence.
11 Furthermore, Lee and See (2004) summarized the factors affecting trust in automation
12 into a three-dimensional model, including performance, process, and purpose. Jian et al.
13 (2000) built a scale system to measure trust using an experimental study that explored the
14 similarities and differences between trust and distrust in automation. Raue et al. (2019)
15 used linear regression ($R^2 = 0.72$) to model interests in using AVs and logistic
16 regression ($R^2 = 0.31$) to model parents' attitudes toward children riding in AVs alone.
17 Both models identified significant factors (e.g., risk perception, benefit perception,
18 negative emotions in manual driving) influencing the dependent variables, but no
19 prediction results were reported. Commonly, trust models are modeled using a linear
20 combination of the input factors, which identify significant factors that influence trust in
21 AVs and other automation systems. However, they did not report prediction results.
22 Therefore, machine learning techniques were proposed in modeling trust in AVs. For
23 example, Liu et al. (2011) investigated the usage of two machine learning models: linear
24 discriminant analysis for feature importance and decision trees for classification for
25 large-scale systems (e.g., product recommendation systems, Internet auction sites) with
26 false rates between 10% and 19%. Guo and Yang (2020) developed a personalized trust
27 prediction model based on the Beta distribution and learned its parameters using
28 Bayesian inference. López and Maag (2015) designed a generic trust model capable of
29 processing various trust features with an SVM technique. On their simulated trust dataset,
30 they obtained 96.61% accuracy. Akash et al. (2018) developed an empirical trust model
31 of object detection in AVs and they used a quadratic discriminant classifier and

1 psychophysiological measurements, such as electroencephalography (EEG) and galvanic
 2 skin response (GSR). Their model’s best accuracy was 78.55%. Such models were able to
 3 predict people’s trust in AVs to a large extent by aggregating numerous factors.
 4 However, the relative importance in predicting trust in AVs tends to be not obvious in
 5 such black-box models. Unlike prior work, we propose a research method that combines
 6 XGBoost and SHAP to help increase the predictability and explainability of trust in AVs,
 7 simultaneously.

8 3. System architecture

9 The proposed system architecture is illustrated in Fig. 1 with the following steps:

- 10 (1) Data Collection: We collected a dataset using an online survey on Amazon
 11 Mechanical Turks (AMTs). The survey was developed in Qualtrics and it was
 12 integrated in AMT to collect participants’ responses.
 13 (2) Data Cleaning: We reviewed the participants’ responses and removed invalid
 14 data.
 15 (3) XGBoost Model Construction: We used a 10-fold cross validation process to
 16 optimize the parameters of XGBoost to train the model.
 17 (4) XGBoost Model Evaluation: To evaluate the performance of the XGBoost model,
 18 we compared it with a list of machine learning models using various performance
 19 metrics, including accuracy, receiver operator characteristics area under the curve
 20 (ROC_AUC), precision, recall, and F1 measure.
 21 (5) SHAP Explanation: To improve the explainability of the XGBoost model, SHAP
 22 was used to explain the model predictions both globally and locally.

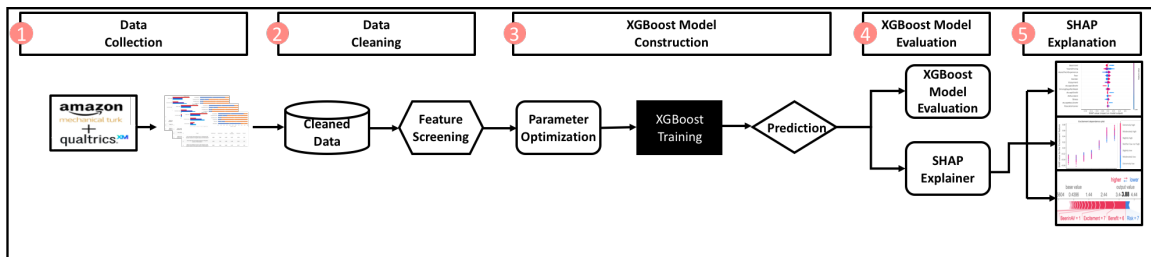


Fig. 1. Flow chart of the proposed system architecture to predict Trust.

1 **4. Method**

2 **4.1. Participants and Apparatus**

3 A total number of 1175 participants located in the United States took part in the online
4 survey using AMTs (Seattle, WA, www.mturk.com/). AMT is a web-based survey
5 company, operated by Amazon Web Services, which has recently become popular in fast
6 data collection (Paolacci et al., 2010). The questionnaire was developed in Qualtrics
7 (Provo, UT, www.qualtrics.com), a web-based software to create surveys. Participants
8 who gave nonsensical answers (i.e., unreasonable driving experience compared to their
9 age, using letters instead of numbers to represent the number of driving years, using the
10 same pattern to answer all the questions, and completing the survey too quickly) were
11 excluded from the study. After the screening, we had a total number of 1054 participants
12 (47.5% females, 52.2% males, and 0.3% others). The age distribution and the education
13 distribution of the participants are shown in Table 1. Participants were compensated with
14 \$0.2 upon completion of the survey. The study was approved by the Institutional Review
15 Board at the University of Michigan.

16 **Table 1.** Age and education distribution of the participants in the study

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

17
18 **4.2. Survey Design**

19 We investigated various factors associated with AVs, including knowledge, experience,
20 feelings, risk and benefit perceptions, and behavioral assessment to predict trust using a
21 survey study. The survey questions were adapted from (Raue et al., 2019; Jian et al.,
22 2000) and are shown in Table 2. Participants’ knowledge about AVs was measured using
23 their eagerness level to adopt a new technology, knowledge level about AVs, and
24 knowledge about AV crashes. Experience questions were related to the experience of
25 using ADAS and the experience of trying AVs. As for Benefit and Risk related questions,
26 participants had to assess how beneficial and risky the AVs were. In regard to the

1 behavioral assessment related questions, participants were asked if they would let a child
2 under 5 years old, between 6 to 12 years old, between 13 to 17 years old, and above 18
3 years old use an AV alone. Since the majority of the public had no experience in AVs
4 yet, we asked them to rate their feelings (i.e., Control, Excitement, Enjoyment, Stress,
5 Fear, and Nervousness) based on their experience in manual driving. Among all the items
6 in the survey, those related to knowledge and experience directly measured participants'
7 initial learned trust while others measured their dispositional trust. We provided
8 abbreviated names for the survey questions to use them throughout the paper as shown in
9 Table 2.

10 **4.3. XGBoost Model Construction**

11 In this study, the XGBoost classifier was selected for predicting trust in AVs (Chen and
12 Guestrin, 2016). The boosting algorithm combines multiple decision trees into a strong
13 ensemble model and reduces the bias by reducing the residual error at each iteration
14 where each decision tree learns from the previous one. This process is done by adjusting
15 the weights of decision trees while iterating the model sequentially. More accurate
16 decision trees are given more weights. XGBoost implements the same boosting technique
17 with an additional regularization term. During the optimization process, an optimal
18 output value for each tree is obtained by iteratively splitting each tree to minimize its
19 objective function.

20 To build a tree, the process follows the exact greedy algorithm where it starts with all the
21 training examples, and then it calculates the split loss reduction or gain for the root of the
22 tree. Once the gain for all the split trees is calculated, the tree with the maximum gain is
23 considered as the optimal split. The gain value should be positive in order for the selected
24 tree to continue growing. After building the trees, pruning is performed to remove the
25 sections with low effect on the classification. Then, an output value is calculated for each
26 leaf which will be used to make predictions. Using these predictions, the same described
27 process is used to build a second tree. The XGBoost algorithm combines both software
28 and hardware optimization abilities, which result in great performance with less
29 computational resources by performing parallel computing.

Table 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	
	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	Yes / No
	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	
	19) How much do you feel excited when you are driving?	Excitement	
	20) How much do you enjoy driving?	Enjoyment	From 1 (extremely low) to 7 (extremely high)
	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)

1 In this research, we removed the highly correlated predictor variables before starting the
2 training process in XGBoost using the Pearson correlation coefficient. The correlation
3 coefficient was high between age and number of driving years (0.88) and between fear
4 and nervousness (0.87). Therefore, age and nervousness were removed. We defined the
5 response variable as a binary one, (i.e., trust = 1 (extremely high, moderately high, and
6 slightly high), sample size = 624, and distrust = 0 (extremely low, moderately low and
7 slightly low), sample size = 430) by converting its 7-point Likert scale. In the next step,
8 we trained the XGBoost classifier with 10-fold cross validation to optimize the accuracy
9 of the prediction using a randomized search for hyperparameters. The learning objective
10 used in this study was reg: logistic regression. After we constructed the model, we
11 compared XGBoost with other machine learning models using various performance
12 metrics, including accuracy, ROC_AUC, precision, recall, and F1 measure. Accuracy is
13 the fraction of corrected prediction samples divided by the total samples. ROC plots the
14 true positive rate against the false positive rate at various threshold settings, and
15 ROC_AUC measures the performance of a classifier in distinguishing between the two
16 classes. Precision is defined as true positive/(true positive + false positive), recall as true
17 positive/(true positive + false negative), and F1 measure as the harmonic mean of
18 precision and recall, i.e., $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$ (Zhou et al., 2017).

19 **4.4. Explaining XGBoost Model Using SHAP**

20 Shapley value is a method from coalitional game theory (Shapley, 1953), in which each
21 player is assigned with payouts depending on their contribution to the total payout when
22 all of them cooperate in a coalition. In our study, in the case of XGBoost model, each
23 feature (i.e., predictor variables in XGBoost) has its fair contribution to the final
24 prediction of trust perception on AVs. Predicting if one participant trusts or distrusts AVs
25 can be considered as a game, and the gain in this game is the actual prediction for this
26 participant minus the average prediction for all the participants' data. For example, if we
27 use three feature-value sets, i.e., Benefit = 7, BeeninAV = 1, and KnowledgeinAVs = 7 to
28 predict trust in AVs, the predicted Trust is 7 and if we use Benefit = 7 and
29 KnowledgeinAVs = 7 to predict trust in AVs, the predicted Trust is 5. Assuming we want
30 to calculate the Sharply value of the feature-value set, BeeninAV = 1, the contribution
31 from the above example is $7 - 5 = 2$ in trust prediction. However, this is only one

1 coalition, we need to repeat the same process for all the possible coalitions and obtain the
 2 average of all the marginal contributions. Mathematically, the Shapley value of a feature-
 3 value set is calculated as follows (Shapley, 1953):

$$4 \quad \varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} (v(S \cup \{i\}) - v(S)), \quad (1)$$

5 where n is the total number of features, S is a subset of any coalition of the features N ,
 6 where the summation extends over all subsets S of N that do not contain feature i , and
 7 $v(S)$ is the contribution of coalition S in predicting trust in our study. The difference
 8 between the trust prediction and the average trust prediction is fairly distributed among
 9 all the feature-value sets in the data. Therefore, it has a solid theory in explaining
 10 machine learning models.

11 One limitation is that when the number of features increases (so is the exponential
 12 number of coalitions), the computation needed will be exponentially expensive.

13 According to game theoretically optimal Shapley values, Lundberg and Lee (2017) and
 14 Lundberg et al. (2020) proposed an efficient method to calculate SHAP values, especially
 15 for tree-based models, such as XGBoost. Therefore, we can use SHAP to explain
 16 XGBoost both globally and locally. Globally, we can study how SHAP values rank the
 17 features based on their importance, how SHAP values change with regard to different
 18 feature-value sets, and how one feature interacts with another. Locally, we can explain
 19 individual predictions. Among them, the interaction effect is defined as the additional
 20 combined feature effect minus individual main feature effects:

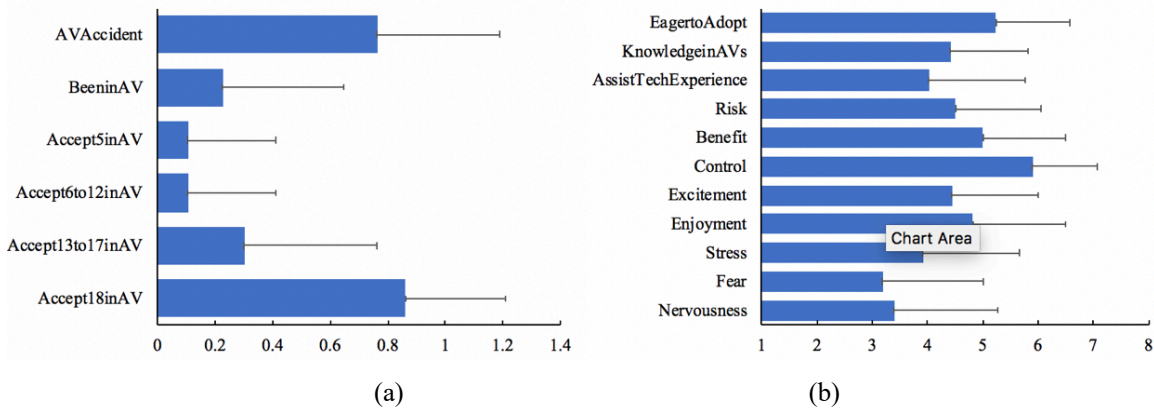
$$21 \quad \varphi_{i,j}(v) = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(n-|S|-2)!}{n!} (v(S \cup \{i,j\}) - v(S \cup \{i\}) - v(S \cup \{j\}) + v(S)), \quad (2)$$

22 Thus, SHAP can produce an n by n interaction matrix and automatically can identify the
 23 strongest interaction effect given one specific feature. In this research, after training the
 24 XGBoost model, SHAP was used to explain the model predictions (Lundberg and Lee,
 25 2017) by calculating the importance of each feature, by evaluating the interaction effects
 26 between the features globally, and by explaining individual predictions locally.

1 **5. Results**

2 **5.1. Survey Results**

3 We calculated participants' mean responses and the standard deviations as shown in Fig.
4 2. The knowledge-related questions indicated that the majority of the participants had a
5 relatively high level of knowledge about AVs — 75.1% had a high level of eagerness to
6 adopt a new technology (i.e., by high, we mean a Likert scale value greater than or equal
7 to 5, moderate refers to a Likert scale value of 4, and low refers to a Likert scale value
8 less than or equal to 3), 51% had a high level of knowledge in AVs, and 76.4% of the
9 participants knew about accidents related to AVs. As for the experience related questions,
10 the majority showed a low level of experience in AVs— 46% of the participants had a
11 high level of experience in ADAS and 77.3% had never been in an AV.



12
13
14 **Fig. 2.** Mean values and standard deviations of the predictor variables. (a) “0” =
15 No, 1 = “Yes”; (b) “1” = Extremely low, “2” = Moderately low, “3” = Slightly
16 low, “4” = Neither low nor high, “5” = Slightly high, “6” = Moderately high, “7”
17 = Extremely high.

18 Furthermore, the majority considered AVs as beneficial (71%), but risky (57%). In regard
19 to behavioral assessment of AVs, 89% of the participants were reluctant to let a child
20 under 5 or between 6 to 12 use an AV alone and 70% were reluctant to let a child
21 between 13 and 17 use an AV alone. However, 86% were willing to let a child above 18
22 use an AV alone. Feelings related questions showed that the majority of the participants
23 reported a high level of control (91%) and a high level of excitement (51%) and
24 enjoyment (64%) while driving. In addition, 58% of the participants had a low level of

1 fear and nervousness of driving, but 44% of the participants considered driving as being
 2 stressful.

3 **5.2. XGBoost Performance**

4 The performance of the XGBoost prediction model, including accuracy, ROC_AUC,
 5 precision, recall, and F1 measure, is shown in Table 3 using a 10-fold cross validation
 6 strategy. In order to compare the performance of XGBoost with other algorithms (see
 7 Table 3), we also performed a 10-fold cross validation strategy. We found that XGBoost
 8 performed the best across almost all the metrics (except precision) among the list of the
 9 machine learning models including logistic regression, decision trees, naive Bayes, linear
 10 SVM, and random forest.

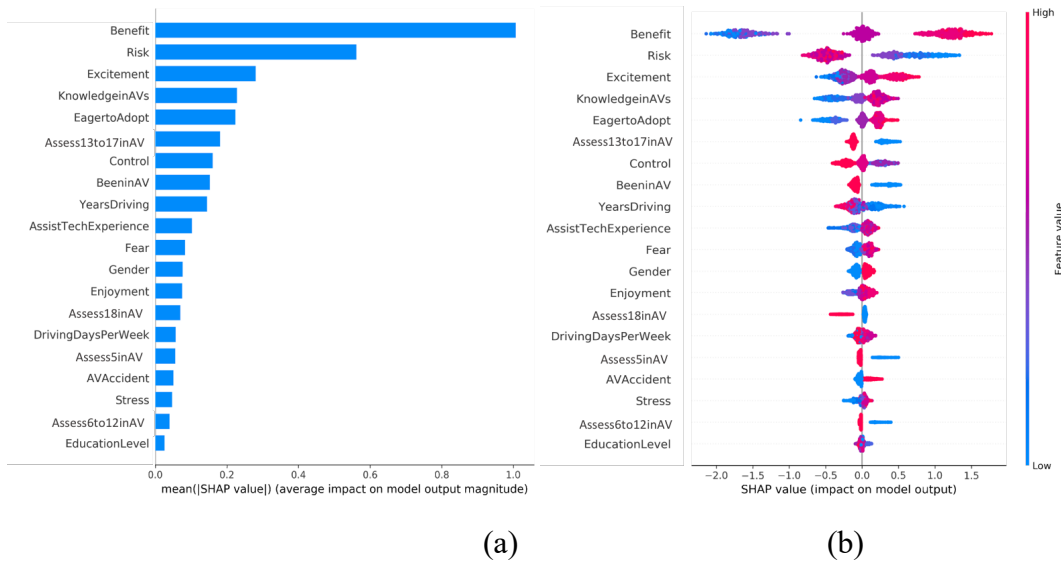
11 **Table 3.** Performance measures comparison between different models

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%

12 **5.3. SHAP Global Explanation**

13 **Importance of Predictor Variables:** To understand the importance of each factor in
 14 predicting trust in AVs, we examined SHAP feature (i.e., predictor variable) importance
 15 and summary plots. The SHAP feature importance plot sorts the features by the mean of
 16 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
 17 number of the samples. The SHAP summary plot also combines feature importance with
 18 feature effect. Note the unit of the SHAP value here is log odds as the objective function
 19 was set as logistic regression in training the XGBoost model. The summary plot lists the
 20 most significant factors in a descending order as illustrated in Fig. 3(a). The top factors
 21 (e.g., Benefit, Risk, Excitement, KnowledgeinAVs, EagertoAdopt) contributed more to
 22 the prediction. To obtain more information about the factors, we also explored the
 23 summary plot in Fig. 3(b). Each data point (i.e., each participant) has three

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



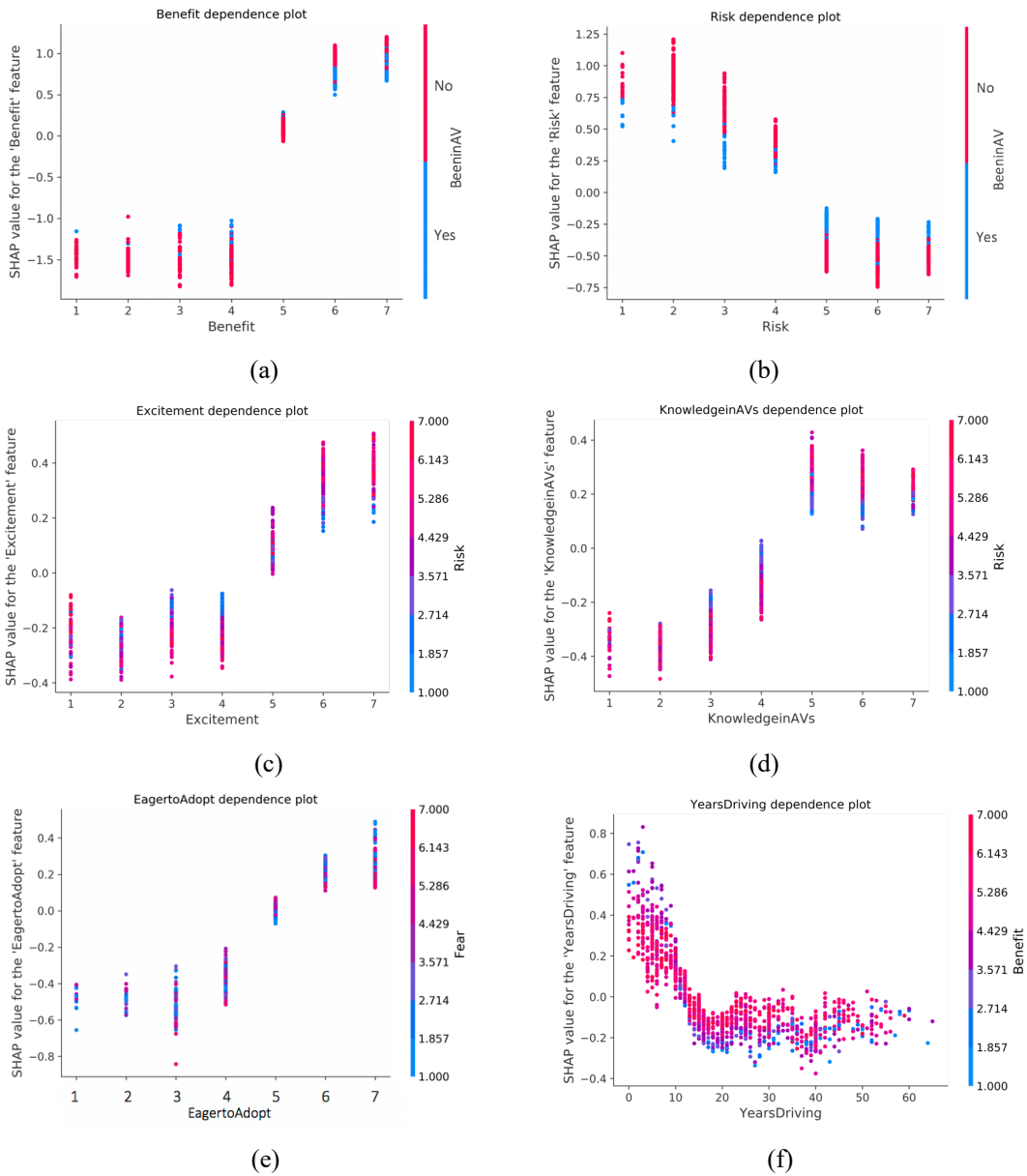
12
 13
 14 **Fig. 3.** (a) SHAP feature importance plots (b) SHAP summary plot.

15 **Dependence Plot:** To further understand the relationship between the predictor variables
 16 and the response variable, we examined their individual SHAP dependence plots which
 17 can capture both, the main effects of individual predictor variables and the interaction
 18 effects between predictor variables. Fig. 4 shows the SHAP dependence plots of the top
 19 five most important factors (i.e., Benefit, Risk, Excitement, KnowledgeinAVs, and
 20 EagertoAdopt) and a continuous variable, i.e., YearsDriving. For instance, to understand
 21 the impact of Benefit on trust as captured by the XGBoost model, the SHAP dependence
 22 plot is shown in Fig. 4(a). The horizontal axis represents the actual values of the Benefit
 23 factor from the dataset, and the vertical axis represents the effect of the factor on the

1 prediction. For the main effect, the plot shows an increasing trend between the factor
2 Benefit and the target trust. It also shows the interaction effect between Benefit and
3 BeeninAV automatically selected by the SHAP model. Out of the participants who
4 scored low on benefits of AVs, those who had experience with AVs trusted AVs more
5 than those who had no experience. On the other hand, out of the participants who scored
6 high on benefits of AVs, those who had experience with AVs trusted AVs less than those
7 who had no experience with AVs. The SHAP dependence plot of Risk is illustrated in
8 Fig. 4(b). We can observe that risks in AVs are negatively correlated with trust in AVs.
9 Meanwhile, among the participants who scored low on risks in using AVs, those who had
10 no experience in AVs trusted AVs more than those with experience. On the other hand,
11 among the participants who scored high on risks in AVs, those who had experience with
12 AVs trusted AVs more than those who did not. The effect of Excitement on trust is
13 illustrated in Fig. 4(c). The higher the excitement about manual driving, the higher the
14 likelihood to trust AVs. And among the participants with a low level of excitement about
15 driving, those who scored high on perceived risks in AVs trusted AVs less than those
16 who scored low on perceived risks. However, among the participants with a high level of
17 excitement about driving, those who scored high on perceived risks in AVs trusted AVs
18 more than those who scored low on perceived risks in AVs. Fig. 4(d) illustrates the effect
19 of KnowledgeinAVs on trust. The increasing slope indicates that the more the
20 Knowledge in AVs, the higher the likelihood to trust AVs. For the participants who rated
21 low in knowledge in AVs, those with low perceived risks in AVs trusted AVs more than
22 those with high perceived risks in AVs. However, when the participants rated high in
23 knowledge in AVs, those with high perceived risks in AVs trusted AVs more than those
24 with low perceived risks in AVs. The increasing slope in Fig. 4(e) shows that the more
25 eager the participants are to adopt a new technology, the higher the likelihood is to trust
26 AVs. Out of the participants who were not eager to adopt a new technology, the
27 interaction effect was not clear. However, out of the participants who were eager to adopt
28 a new technology, those being not fearful of driving trusted AVs more than those being
29 fearful of driving. In Fig. 4(f), we see a decreasing slope which illustrates that people
30 with more experience in driving are less likely to trust AVs. For the participants with

1 driving experience between 10 and 40 years, those who reported a high level of perceived
 2 benefits trusted AVs more than those who reported a low level of perceived benefits.

3
 4

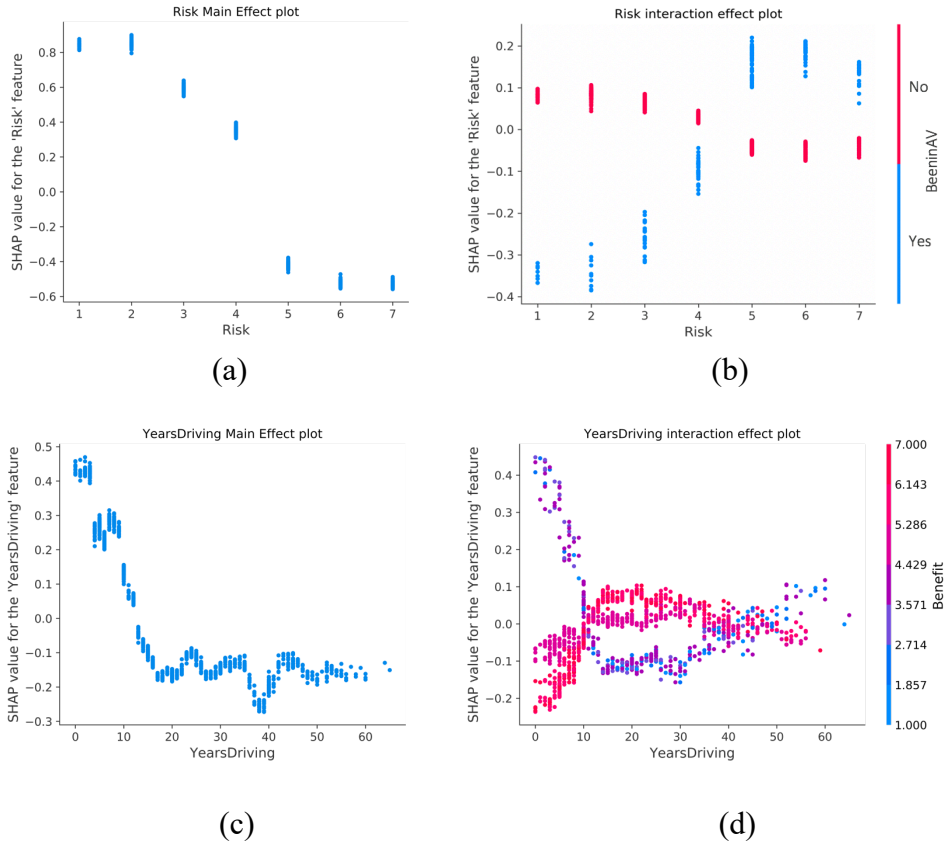


5
 6

7
 8

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

13



1
2

3
4

5 **Fig. 5.** SHAP main effects and interaction effects derived from SHAP dependence plots.
6 “1” = Extremely low, “2” = Moderately low, “3” = Slightly low, “4” = Neither low nor
7 high, “5” = Slightly high, “6” = Moderately high, “7” = Extremely high.

8 **Main Effects and Interaction Effects:** The SHAP dependence plot has rich information,
9 which incorporates both main effects of individual predictor variables and interaction
10 effects between two predictor variables. The interaction effects are demonstrated by the
11 vertical dispersion as shown in Fig. 4. Such interaction shows the effect of the two
12 predictor variables on the response variable at the same time. We can also separate the
13 main effects and interaction effects in individual plots. Take the Risk SHAP dependence
14 plot in Fig. 4(b) as an example. Its main effect and interaction effect with BeeninAV are
15 shown in Fig. 5(a) and Fig. 5(b). There is little vertical dispersion in the main effect. The
16 interaction effect is also more apparent suggesting that at lower Risk levels, participants
17 who experienced AVs trusted AVs less than those who did not experience AVs.
18 However, at higher Risk levels, participants who experienced AVs trusted AVs more than
19 those who did not experience AVs. Take the YearsDriving as another example. Its main

1 effect and interaction effect with Benefit are shown in Fig. 5(c) and Fig. 5(d). Also, less
 2 vertical dispersion is observed in the main effect plot, and the interaction effect tends to
 3 be more apparent. That is, only when YearsDriving is larger than 10 and smaller than 40,
 4 more Benefits lead to a stronger likelihood to trust AVs.

5 In Table 4, we presented the sum of the main effects (i.e., $\sum_{j=1}^M |\phi_{ij}(v)|$, where M is the
 6 total number of the samples) and selected interaction effects of the six predictor variables
 7 corresponding to Fig. 4. The larger the magnitudes of the main/interaction effects, the
 8 more important they are to predict trust. Furthermore, we also calculated the correlation
 9 coefficients between the selected predictor variables and their SHAP values and between
 10 the selected predictor variables and the response variable, i.e., trust. Although all the
 11 correlations are significant, the correlations with SHAP values are stronger, indicating
 12 that XGBoost tends to capture the correlations better than linear models.

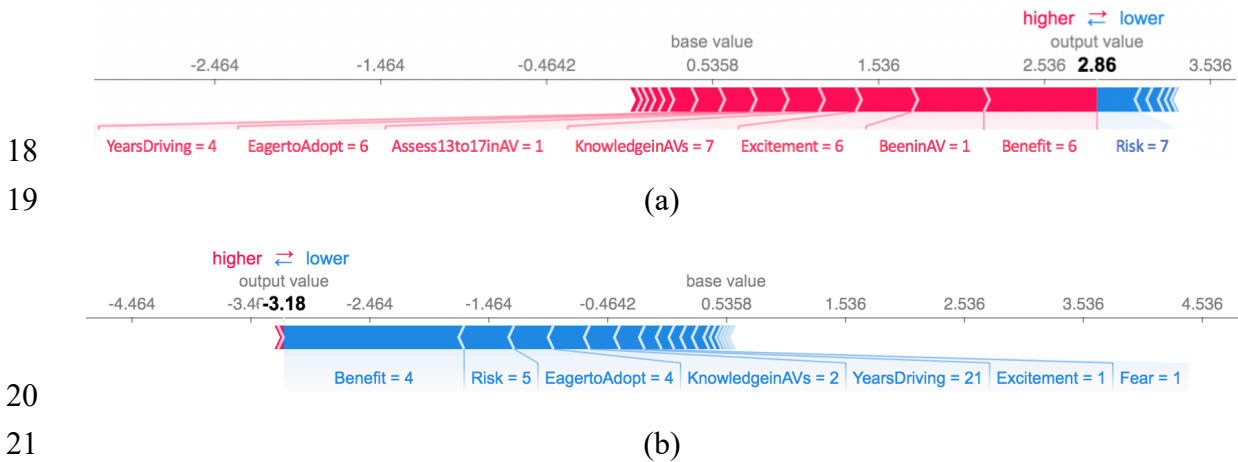
13
 14 **Table 4.** Rich information obtained from SHAP dependent plots for selected predictor
 15 variables

Predictor Variables	Main effect	Selected interaction effect	Correlation with SHAP values	Correlation with Trust
Benefit	945.94	:BeeninAV: 22.13	0.89	0.61
Risk	543.47	:BeeninAV: 41.67	-0.90	-0.37
Excitement	233.42	:Risk: 31.39	0.86	0.25
KnowledgeinAVs	265.51	:Risk: 24.51	0.87	0.41
EagertoAdopt	234.70	:Fear: 5.96	0.92	0.42
YearsDriving	190.59	:Benefit: 39.02	-0.69	-0.24

16 The p values of all the correlation coefficients in the table are smaller than 0.001

1 **5.4. SHAP Local Explanations**

2 In order to show how SHAP explains individual cases, we tested it on two randomly
 3 selected observations as illustrated in Fig. 6. The plots show the different factors
 4 contributing to pushing the output value from the base value which represents the average
 5 model output over the training dataset. The base value is defined as the mean prediction
 6 value (Lundberg et al., 2018), which is 0.5358 in our case. Factors pushing the SHAP
 7 value (i.e., log odds) larger are shown in red while those pushing the SHAP value smaller
 8 are shown in blue. In Fig. 6(a), the model produced a large SHAP value in predicting
 9 trust which was consistent with the ground truth (i.e., trust) because the participant
 10 perceived the AV with a high level of Benefits (i.e., 6), BeeninAV = Yes, a high level of
 11 Excitement (i.e., 6), a high level of KnowledgeinAVs (i.e., 7), Assess13to17inAV = Yes,
 12 a high level of EagertoAdopt (i.e., 6), YearsDriving (i.e., 4), even though the participant
 13 perceived the AV with a high level of Risk (i.e., 7). In Fig. 6(b), the model produced a
 14 small SHAP value, which was consistent with the ground truth (i.e., distrust) mainly due
 15 to a neutral level of Benefit, a high level of Risk (i.e., 5), a neutral level of EagertoAdopt
 16 (i.e., 4), a low level of KnowledgeinAVs (i.e., 2), 21 YearsDriving, a low level of
 17 Excitement (i.e., 1), and a low level of Fear (i.e., 1).



22 **Fig. 6.** SHAP individual explanations of trust prediction for randomly selected
 23 participants with (a) ground truth = trust and (b) ground truth = distrust.
 24

1 **6. Discussion**

2 **6.1. Predictability and Explainability**

3 XGBoost is an efficient and easy to use algorithm for tabular data classification and
4 delivers high performance and accuracy as compared to other algorithms (Chen and
5 Guestrin, 2016). In this research, we used XGBoost to predict people's trust in AVs with
6 superior performance. Compared to other machine learning models, XGBoost performed
7 the best among various metrics, including accuracy, ROC_AUC, recall, and F1 measure
8 (see Table 3). The model converged within 60 iterations in our experiment and proved to
9 be a feasible solution to predict trust in AVs.

10 In order to improve the explainability of the XGBoost model, we used SHAP explainer
11 which offers a high level of model interpretability (Lundberg and Lee, 2017). SHAP has
12 a fast implementation for tree-based models (e.g., XGBoost), which overcomes the
13 biggest barrier (i.e., slow computation) for adoption of Shapley values. On top of the
14 advantage of fast implementation, SHAP provides another two advantages including
15 global and local interpretability. The global interpretability is represented by the
16 contribution of the SHAP values in the model predictive decision. It represents the
17 negative and positive effects of the most important factors on the model prediction as
18 shown in Fig. 3. Such global interpretability is similar to the feature effect plot in linear
19 regression models. Furthermore, the model is able to show both main effects of
20 individual predictor variables and interaction effects between two predictor variables
21 on trust, indicating how they influence the prediction results as evidenced in Fig. 4, Fig.
22 5, and Table 4. As for the local interpretability, SHAP enables us to explain the
23 prediction of each observation since each one gets its own set of SHAP values as
24 illustrated in Fig. 6. With the local and global interpretability comes the power of SHAP
25 in providing a high level of model explainability.

26 **6.2. Important Factors in Predicting Trust**

27 Compared to linear regression models, our method uncovered the factor importance in
28 predicting trust using the SHAP feature importance plots and the SHAP summary plot as
29 shown in Fig. 3. Among all the predictor variables, the Benefit factor ranked the most

1 important and was positively correlated with trust, consistent with previous research
2 (Choi and Ji, 2015; Bearth and Siegrist, 2016). Furthermore, we found an interaction
3 effect between Benefit and BeeninAV (see Fig. 4(a)). Even when the participants
4 perceived AVs with low benefits, their interaction with AVs could potentially improve
5 their trust in them. This was consistent with Brell et al. (2019), which showed that the
6 experience with AVs significantly increased the perception of the benefits in AVs.
7 The second most important factor was risk (Fig. 3). In line with prior studies (Numan,
8 1998; Kim et al., 2008; Pavlou, 2003), our results showed that an increase in risk led to a
9 decrease in trust. Risk was found to interact with BeeninAV (Fig. 4(b)). When the
10 participants viewed AVs to be risky, experience with AV could potentially improve their
11 trust in AV. This was also in concordance to previous research (Brell et al., 2019), which
12 showed a decrease in risk perception in AVs with the increase of experience in AVs.
13 Therefore, it is important that automotive manufacturers give more chances for the public
14 (especially for those who perceive AVs with no benefits or high risks) to test AVs in
15 order to improve their trust in AVs.

16 While both the third and fourth most important factors, i.e., Excitement and
17 KnowledgeinAVs were positively correlated with trust in AVs. Risk was found to
18 interact with Excitement (Fig. 4(c)) and KnowledgeinAVs (Fig. 4(d)). When the
19 participants were not very excited about manual driving, they tended to trust the AVs
20 more if the risk was low. Silberg et al. (2013) found that people who were less passionate
21 about driving were more likely to lean toward using AVs if it was safe. When the
22 participants were excited about manual driving, they trusted the AV more even if the risk
23 was higher. Such trust, however, could be a type of over-trust associated with strong
24 emotions, such as excitement. For example, Dingus et al. (2016) argued that excited or
25 angry drivers were more likely to take risky driving even in highly automated driving. An
26 increase in KnowledgeinAVs increased participants' trust in AVs (Fig. 4(d)) which was
27 consistent with previous studies such as (Khastgir et al., 2018). However, it seemed
28 counter-intuitive that those who rated AVs as risky trusted AVs more than those who
29 rated AVs as not risky when the participants scored high on knowledge in AVs. To
30 investigate the obtained results, we found that the percentage of participants who scored
31 high on both KnowledgeinAVs and Risk was 27.9. In addition, out of those participants,

1 77.5 % considered AVs as beneficial. Thus, this result might be explained by the finding
2 that the degree of knowledge in AVs affected the perception of balance between the risks
3 and trust in AVs as Schmidt (2004) argued that the more one knew about the risks in an
4 automation system, the higher the chances to accept it. In other words, these participants
5 believed that the risky situations associated with AVs might be avoided by a better
6 understanding of how to deal with such situations, such as the takeover transition period
7 in SAE Level 3 AVs (Zhou, Yang and Zhang, 2020; Na, Yang and Zhou, 2020).
8 Moreover, the belief of the benefits brought from AVs might also make them trust AVs
9 more.

10 The EagertoAdopt factor was ranked number 5, and an increase in eagerness to adopt a
11 technology increased the chances of trusting AVs which was in line with previous
12 research (Edmonds, 2019; Raue et al., 2019) (see Fig. 4(e)). We also found that Fear
13 affected the impact of EagertoAdopt on trust—at a high level of eagerness to adopt a new
14 technology, a low level of fear in manual driving increased the chances of trusting AVs.
15 Fear, which is an important factor in technology adoption, was shown to shape
16 judgements, choices, and perception of risks (Lerner and Keltner, 2001). According to
17 Shoemaker (2018), fearless driving was associated with no fear of change, thus leading to
18 an eagerness of technology adoption.

19 Other factors involved in the study were less important compared to the ones listed
20 above. Although Assess13to17inAV was ranked number 6, it was surprising to see that
21 Assess5inAV and Assess 6to12inAV were less important in predicting trust in AVs.
22 Intuitive, without trust in AVs, a parent would not let his/her children be in an AV.
23 However, in our survey, we did not specify if they were the participants' children.
24 Further research is needed to address this issue. Gender, age (years of driving), and
25 education level were also found to be less important. However, as seen in Fig. 4(f), we
26 found that trust was shown to decrease with an increase in the number of driving years.
27 Furthermore, Benefit affected the impact of DrivingYears on trust—for larger than 10
28 years and smaller than 40 years of driving experience, a high level of benefits increased
29 trust in AVs. In line with previous research, old people showed more concerns about
30 trusting AVs despite its benefits in maintaining their mobility (Schoettle and Sivak,
31 2016).

1 As a summary, the measured trust is based on dispositional trust and initial learned trust
2 (see Hoff and Bashir, 2015). The dispositional trust shows participants' overall tendency
3 without any context of AVs and the initial learned trust is dependent on their previous
4 knowledge or past experience (e.g., news reports on AV accidents) prior to interacting
5 with AVs. This is because the majority of the participants (i.e., 77.3%) had no chance to
6 interact with AVs and there was no interaction between the participants and AVs during
7 this study. However, the dispositional trust and the initial learned trust measured in our
8 paper are the baseline to form people's trust in AVs. Prior to any interaction with AVs,
9 people have an inherent level of dispositional trust which is one of the major factors that
10 influences people's purchase or use of AVs. Individual differences, such as age, gender,
11 educational levels, as well as their learned knowledge about and experience in AVs
12 shaped their perceived risks in and benefits of AVs, which in terms influence their
13 dispositional and initial learned trust. Between these two types of trust measured in the
14 survey, we found that the variables related to dispositional trust were more important and
15 predictive than those related to initial learned trust as shown in Fig. 3(a). Nevertheless,
16 unlike previous studies, the most important contribution of this study was proposing a
17 trust prediction model with explainability to understand participants' trust in AVs.
18 Automotive manufacturers can potentially make use of the relationships between these
19 important factors and their trust to improve acceptance and adoption of AVs by providing
20 training, spreading the benefits of AVs, explaining the possible risks, improving the
21 design of the system, and creating appropriate emotional responses to AVs.

22 **6.3. Limitations**

23 First, due to the cross-sectional study design, we cannot examine how people's opinions
24 change over time. Therefore, we only measured participants' trust in AVs in a snapshot.
25 Also, as the majority of the participants had little prior experience with AVs, the trust is
26 primarily based on their dispositional and initial learned trust. Longitudinal studies are
27 needed to understand the dynamic trust relationships between users and AVs when they
28 have chances to interact with AVs over time (Ekman et al., 2018). Further research
29 should also be conducted to assess participants' dispositional, situational, and learned
30 trust (see Hoff and Bashir, 2015) at a finer granularity, by querying participants' trust in

1 AVs over time (Ruijten et al. 2018). Second, it was difficult for us to make sure the
2 superior quality of the survey data from AMT. In this research, we used various
3 techniques to overcome that, including shorter surveys, removing invalid data by
4 examining their survey completion time and data patterns. However, quality can be
5 affected by the compensation rate (Buhrmester et al., 2011) and running the screening
6 procedures mentioned above might not be enough to ensure a high quality of responses.
7 Third, our survey was quantitative without any qualitative data to explain our prediction
8 model. It would be also important to verify such explanations using qualitative data from
9 the participants themselves with open-ended questions.

10 **7. Conclusion**

11 In this paper, we predicted dispositional and initial learned trust in AVs with high
12 accuracy and explainability. We conducted an online survey to collect a range of
13 variables that were related to participants' trust in AVs. The survey data were then used
14 to train and test an XGBoost model. In order to explain the XGBoost prediction results,
15 SHAP was used to identify the most important predictor variables, to examine main and
16 interaction effects, and to illustrate individual explanation cases. Compared with previous
17 trust predictions models, our proposed method combines the benefits of XGBoost and
18 SHAP with good explainability and predictability of the trust model.

19 **References**

- 20 Abe, G., Sato, K., & Itoh, M. (2017). Driver Trust in Automated Driving Systems: The Case of Overtaking
21 and Passing. *IEEE Transactions on Human-Machine Systems, PP*, 1–10.
22 <https://doi.org/10.1109/THMS.2017.2781619>
- 23 Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial
24 Intelligence (XAI). *IEEE Access, 6*, 52138–52160.
25 <https://doi.org/10.1109/ACCESS.2018.2870052>
- 26 Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A Classification Model for Sensing Human Trust in
27 Machines Using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems, 8(4)*, 1–20.
28 <https://doi.org/10.1145/3132743>
- 29 Ayoub, J., Zhou, F., Bao, S., & Yang, X. J. (2019, September). *From Manual Driving to Automated*
30 *Driving: A Review of 10 Years of AutoUI*. In: Proceedings of the 11th International Conference on
31 Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '19),
32 ACM, New York, NY, USA. <https://doi.org/10.1145/3342197.3344529>

- 1 Ayoub, J., Mason, B., Morse, K., Kirchner, A., Tumanyan, N., & Zhou, F. (2020, April). Otto: An
2 Autonomous School Bus System for Parents and Children. In *Extended Abstracts of the 2020 CHI*
3 *Conference on Human Factors in Computing Systems* (pp. 1-7).
- 4 Bansal, P., Kockelman, K. M., & Singh, A. (2016). Assessing public opinions of and interest in new
5 vehicle technologies: An Austin perspective. *Transportation Research Part C: Emerging*
6 *Technologies*, 67, 1–14. <https://doi.org/10.1016/j.trc.2016.01.019>
- 7 Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is Stronger than Good.
8 *Review of General Psychology*, 5(4), 323–370. <https://doi.org/10.1037/1089-2680.5.4.323>
- 9 Bearth, A., & Siegrist, M. (2016). Are risk or benefit perceptions more important for public acceptance of
10 innovative food technologies: A meta-analysis. *Trends in Food Science & Technology*, 49, 14–23.
11 <https://doi.org/10.1016/j.tifs.2016.01.003>
- 12 Brell, T., Philipsen, R., & Ziefle, M. (2019). sCARY! Risk Perceptions in Autonomous Driving: The
13 Influence of Experience on Perceived Benefits and Barriers. *Risk Analysis*, 39(2), 342–357.
14 <https://doi.org/10.1111/risa.13190>
- 15 Buhrmester, M. D., Kwang, T. N., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A New Source of
16 Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science : A Journal of the*
17 *Association for Psychological Science*. <https://doi.org/10.1177/1745691610393980>
- 18 Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd*
19 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
20 <https://doi.org/10.1145/2939672.2939785>
- 21 Cho, J.-H., Chan, K., & Adali, S. (2015). A Survey on Trust Modeling. *ACM Computing Surveys*, 48, 1–40.
22 <https://doi.org/10.1145/2815595>
- 23 Choi, J., & Ji, Y. G. (2015). Investigating the Importance of Trust on Adopting an Autonomous Vehicle.
24 *International Journal of Human-Computer Interaction*, 31, 150709133142005.
25 <https://doi.org/10.1080/10447318.2015.1070549>
- 26 Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver
27 crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the*
28 *National Academy of Sciences*, 113(10), 2636–2641. <https://doi.org/10.1073/pnas.1513271113>
- 29 Doney, P. M., Cannon, J. P., & Mullen, M. R. (1998). Understanding the Influence of National Culture on
30 the Development of Trust. *The Academy of Management Review*, 23(3), 601–620. JSTOR.
31 <https://doi.org/10.2307/259297>
- 32 Du, N., Yang, X. J., & Zhou, F. (2020). Psychophysiological responses to takeover requests in
33 conditionally automated driving. *Accident Analysis & Prevention*, 148, 105804.
- 34 Du, N., Zhou, F., Pulver, E. M., Tilbury, D. M., Robert, L. P., Pradhan, A. K., & Yang, X. J. (2020).
35 Examining the effects of emotional valence and arousal on takeover performance in conditionally
36 automated driving. *Transportation research part C: emerging technologies*, 112, 78-87.
- 37 Edmonds, E. (2019, March 14). *Three in Four Americans Remain Afraid of Fully Self-Driving Vehicles*.

1 AAA NewsRoom. <https://newsroom.aaa.com/2019/03/americans-fear-self-driving-cars-survey/>

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

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

8 Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: Opportunities,
9 barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77,
10 167–181. <https://doi.org/10.1016/j.tra.2015.04.003>

11 Guo, Y., & Yang, X.J. Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian
12 Inference Approach. *International Journal of Social Robotics* (2020).
13 <https://doi.org/10.1007/s12369-020-00703-3>

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

17 Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That
18 Influence Trust. *Human Factors*, 57(3) 407–434. <https://doi.org/10.1177/0018720814547570>.

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

21 Power. (2012). *Vehicle Owners Show Willingness to Spend on Automotive Infotainment Features.*
22 [https://www.prnewswire.com/news-releases/jd-power-and-associates-reports-vehicle-owners-
23 show-willingness-to-spend-on-automotive-infotainment-features-149088105.html](https://www.prnewswire.com/news-releases/jd-power-and-associates-reports-vehicle-owners-show-willingness-to-spend-on-automotive-infotainment-features-149088105.html)

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

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

31 Johnson, T. (2012, July 25). *Enhancing Safety Through Automation.* In *Society of Automotive Engineers
32 Govt Industry Meeting, Automation and Connected Vehicle Safety, NHTSA.*
33 <http://onlinepubs.trb.org/onlinepubs/conferences/2012/Automation/presentations/Johnson.pdf>

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

37 Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in

- 1 electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support*
2 *Systems*, 44(2), 544–564. <https://doi.org/10.1016/j.dss.2007.07.001>
- 3 Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in
4 automation. *Bagnara S., Tartaglia R., Albolino S., Alexander T., Fujita Y. (Eds) Proceedings of*
5 *the 20th Congress of the International Ergonomics Association (IEA 2018)*. IEA 2018, vol 823.
6 https://doi.org/10.1007/978-3-319-96074-6_2
- 7 Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors:*
8 *The Journal of the Human Factors and Ergonomics Society.*, 31.
9 <https://doi.org/10.1518/hfes.46.1.50.30392>.
- 10 Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*,
11 81(1), 146–159. <https://doi.org/10.1037//0022-3514.81.1.146>
- 12 Lewicki, R., & Brinsfield, C. (2011). Framing trust: Trust as a heuristic. *Framing Matters: Perspectives on*
13 *Negotiation Research and Practice in Communication*, 110–135.
- 14 Lewis, J. D., & Weigert, A. (1985). Trust as a Social Reality. *Social Forces; a Scientific Medium of Social*
15 *Study and Interpretation*, 63(4), Pages 967–985.
- 16 Li, M., Holthausen, B. E., Stuck, R. E., & Walker, B. N. (2019). No Risk No Trust: Investigating Perceived
17 Risk in Highly Automated Driving. *Proceedings of the 11th International Conference on*
18 *Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '19*, 177–185.
19 <https://doi.org/10.1145/3342197.3344525>
- 20 Liu, X., Tredan, G., & Datta, A. (2011). A Generic Trust Framework For Large-Scale Open Systems Using
21 Machine Learning. *Computing Research Repository - CORR*, 30.
22 <https://doi.org/10.1111/coin.12022>
- 23 López, J., & Maag, S. (2015). Towards a Generic Trust Management Framework Using a Machine-
24 Learning-Based Trust Model. *2015 IEEE Trustcom/BigDataSE/ISPA*, 1, 1343–1348.
25 <https://doi.org/10.1109/Trustcom.2015.528>
- 26 Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J.,
27 Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable
28 AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. [https://doi.org/10.1038/s42256-019-0138-](https://doi.org/10.1038/s42256-019-0138-9)
29 9
- 30 Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). Consistent Individualized Feature Attribution for Tree
31 Ensembles. *ArXiv:1802.03888 [Cs, Stat]*. <http://arxiv.org/abs/1802.03888>
- 32 Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of*
33 *the 31st International Conference on Neural Information Processing Systems*, 4768–4777.
- 34 Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-
35 W., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the
36 prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10), 749–760.
37 <https://doi.org/10.1038/s41551-018-0304-0>

- 1 Menon, N. (2015). Consumer Perception and Anticipated Adoption of Autonomous Vehicle Technology:
2 Results from Multi-Population Surveys. *Graduate Theses and Dissertations*.
3 <https://scholarcommons.usf.edu/etd/5992>
- 4 Menon, N., Pinjari, A., Zhang, Y., & Zou, L. (2016, January 1). *Consumer Perception and Intended*
5 *Adoption of Autonomous Vehicle Technology – Findings from a University Population Survey*.
6 <https://trid.trb.org/view/1394249>.
- 7 Miller, D. J. E., & Perkins, L. (2010). *Development of Metrics for Trust in Automation* (p. 18). AIR
8 FORCE RESEARCH LAB WRIGHT-PATTERSON AFB OH SENSORS DIRECTORATE.
9 <https://apps.dtic.mil/docs/citations/ADA525259>
- 10 Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral Measurement
11 of Trust in Automation: The Trust Fall. *Proceedings of the Human Factors and Ergonomics*
12 *Society Annual Meeting, 60*, 1849–1853. <https://doi.org/10.1177/1541931213601422>
- 13 NHTSA2010. (2010). *Traffic Safety Facts 2010 A Compilation of Motor Vehicle Crash Data from the*
14 *Fatality Analysis Reporting System and the General Estimates System*. National Highway Traffic
15 Safety Administration National Center for Statistics and Analysis U.S. Department of
16 Transportation. <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811659>
- 17 Numan, J. H. (1998). *Knowledge-based systems as companions: Trust, human computer interaction and*
18 *complex systems*. Undefined. [https://www.semanticscholar.org/paper/Knowledge-based-systems-](https://www.semanticscholar.org/paper/Knowledge-based-systems-as-companions%3A-Trust%2C-human-Numan/afb2b16ea898a8fd5ec603a38e69c1d742e75e35)
19 [as-companions%3A-Trust%2C-human-Numan/afb2b16ea898a8fd5ec603a38e69c1d742e75e35](https://www.semanticscholar.org/paper/Knowledge-based-systems-as-companions%3A-Trust%2C-human-Numan/afb2b16ea898a8fd5ec603a38e69c1d742e75e35)
- 20 Paden, B., Cap, M., Yong, S. Z., Yershov, D., & Frazzoli, E. (2016). A Survey of Motion Planning and
21 Control Techniques for Self-driving Urban Vehicles. *ArXiv:1604.07446 [Cs]*.
22 <http://arxiv.org/abs/1604.07446>
- 23 Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems.
24 *Communications of the ACM*, 51–55.
- 25 Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human*
26 *Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), 230–253.
27 <https://doi.org/10.1518/001872097778543886>
- 28 Pavlou, P. A. (2003). *Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the*
29 *Technology Acceptance Model* (SSRN Scholarly Paper ID 2742286). Social Science Research
30 Network. <https://papers.ssrn.com/abstract=2742286>
- 31 Peters, E., Västfjäll, D., Gärling, T., & Slovic, P. (2006). Affect and decision making: A “hot” topic.
32 *Journal of Behavioral Decision Making*, 19(2), 79–85. <https://doi.org/10.1002/bdm.528>
- 33 Rajaonah, B., Tricot, N., Anceaux, F., & Millot, P. (2008). The role of intervening variables in driver–ACC
34 cooperation. *International Journal of Human-Computer Studies*, 66(3), 185–197.
- 35 Raue, M., D’Ambrosio, L. A., Ward, C., Lee, C., Jacquillat, C., & Coughlin, J. F. (2019). The Influence of
36 Feelings While Driving Regular Cars on the Perception and Acceptance of Self-Driving Cars:
37 Feelings and Self-Driving Cars. *Risk Analysis*, 39(2), 358–374. <https://doi.org/10.1111/risa.13267>

- 1 Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and
2 Use Interpretable Models Instead. *ArXiv:1811.10154 [Cs, Stat]*. <http://arxiv.org/abs/1811.10154>
- 3 Rudin-Brown, C. M., & Parker, H. A. (2004). Behavioural adaptation to adaptive cruise control (ACC):
4 Implications for preventive strategies. *Transportation Research Part F: Traffic Psychology and*
5 *Behaviour*, 7(2), 59–76. <https://doi.org/10.1016/j.trf.2004.02.001>
- 6 Ruijten, P. A. M., Terken, J. M. B., & Chandramouli, S. N. (2018). Enhancing Trust in Autonomous
7 Vehicles through Intelligent User Interfaces That Mimic Human Behavior. *Multimodal*
8 *Technologies and Interaction*, 2(4), 62. <https://doi.org/10.3390/mti2040062>
- 9 Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk.
10 *Judgment and Decision Making*, 5(5),
11 9. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1626226
- 12 SAE. (2018, June 15). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-*
13 *Road Motor Vehicles*. https://saemobilus.sae.org/content/j3016_201806
- 14 Schmidt, Markus. 2004. Investigating Risk Perception: A Short Introduction. *Loss of Agro-Biodiversity in*
15 *Vavilov Centers, with a Special Focus of Genetically Modified Organisms (GMOs)*, Edited by:
16 *Schmidt, M. , Ph. D. Thesis, Vienna*.
17 http://faculty.mercer.edu/butler_aj/documents/Intro_risk_perception_Schmidt_000.pdf.
- 18 Schoettle, B., & Sivak, M. (2016). *Motorists' Preferences for Different Levels of Vehicle Automation: 2016*
19 (SWT-2016-8). Article SWT-2016-8. <https://trid.trb.org/view/1480408>
- 20 Shapley, L. S. (1953). 17. *A Value for n-Person Games. Contributions to the Theory of Games (AM-28)*,
21 *Volume II*. <https://doi.org/10.1515/9781400881970-018>
- 22 Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving
23 vehicles. *Nature Human Behaviour*, 1. <https://doi.org/10.1038/s41562-017-0202-6>
- 24 Shoemaker, A. (2018). *Fear of Change: Autonomous Vehicle Technology and the Automobile as a Cultural*
25 *Artifact* [Scripps College]. https://scholarship.claremont.edu/scripps_theses/1192
- 26 Silberg, G., Mitch, M., Kevin, E., Deepak, S., Michael, C., Hugh, F., & Vivek, S. (2013). *Self-Driving*
27 *Cars: Are We Ready?* [https://assets.kpmg/content/dam/kpmg/pdf/2013/10/self-driving-cars-are-](https://assets.kpmg/content/dam/kpmg/pdf/2013/10/self-driving-cars-are-we-ready.pdf)
28 [we-ready.pdf](https://assets.kpmg/content/dam/kpmg/pdf/2013/10/self-driving-cars-are-we-ready.pdf)
- 29 Walker, G. H., Stanton, N. A., & Salmon, P. (2016). Trust in vehicle technology. *International Journal of*
30 *Vehicle Design*, 70(2), 157. <https://doi.org/10.1504/IJVD.2016.074419>
- 31 Zhou, F., Lei, B., Liu, Y., & Jiao, R. J. (2017). Affective parameter shaping in user experience prospect
32 evaluation based on hierarchical Bayesian estimation. *Expert Systems with Applications*, 78, 1–15.
33 <https://doi.org/10.1016/j.eswa.2017.02.003>
- 34 Zhou, F., Yang, X. J., & Zhang, X. (2020). Takeover transition in autonomous vehicles: a YouTube
35 study. *International Journal of Human–Computer Interaction*, 36(3), 295-306.
- 36 Zmud, J., N.Sener, I., & Wagner, J. (2016). *Consumer Acceptance and Travel Behavior Impacts of*
37 *Automated Vehicles*. <https://static.tti.tamu.edu/tti.tamu.edu/documents/PRC-15-49-F.pdf>