

Clustering Human Trust Dynamics for Customized Real-time Prediction

Jundi Liu¹, Kumar Akash², Teruhisa Misu² and Xingwei Wu²

Abstract—Trust calibration is necessary to ensure appropriate user acceptance in advanced automation technologies. A significant challenge to achieve trust calibration is to quantitatively estimate human trust in real-time. Although multiple trust models exist, these models have limited predictive performance partly due to individual differences in trust dynamics. A personalized model for each person can address this issue, but it requires a significant amount of data for each user. We present a methodology to develop customized model by clustering humans based on their trust dynamics. The clustering-based method addresses the individual differences in trust dynamics while requiring significantly less data than personalized model. We show that our clustering-based customized models not only outperform the general model based on entire population, but also outperform simple demographic factor-based customized models. Specifically, we propose that two models based on “confident” and “skeptical” group of participants, respectively, can represent the trust behavior of the population. The “confident” participants, as compared to the “skeptical” participants, have higher initial trust levels, lose trust slower when they encounter low reliability operations, and have higher trust levels during trust-repair after the low reliability operations. In summary, clustering-based customized models improve trust prediction performance for further trust calibration considerations.

I. INTRODUCTION

Vehicle automation technologies have significant benefits for society, including dramatic decreases in car crashes, injuries and deaths, increased mobility, increased road efficiency, and better utilization of parking and lands [1], [2]. Besides the advantages for society, they can also improve the driving experience and comfortableness of operations [3], [4]. Even though research has shown substantial benefits of vehicle automation technologies, their acceptance does not seem to keep up with the fast-growing market penetration. One of the most widely adopted methods to solve the acceptance issue is to calibrate trust in these technologies to the appropriate level since trust calibration is essential to accept and rely on vehicle automation [5]. Additionally, misuse of the system due to overtrust should be avoided [6]. Many studies have investigated the possible solutions to calibrate trust. These include paradigms that anticipate human behaviors—such as trust—and inform humans to make optimal choices [7], [8], [9]. However, a primary challenge for such an approach is quantitatively predicting human trust in real-time.

Most current studies use questionnaires to measure self-reports of drivers’ trust levels before, during, or after the

interaction with the automated systems [10], [11], [12]. However, it is challenging to obtain self-reports repeatedly without interrupting the task. As an alternative, recent works have developed dynamic models to capture human trust and estimate it in real-time [13], [14], [15], [9]. There are two approaches for developing such models: a general model for the whole population and a personalized model for each individual to account for individual differences [16]. A general trust model ignores individual differences but can be trained using limited data. On the other hand, a completely personalized model designed for each person requires a significant amount of data for each new user. Such a personalized model may be applicable for some small-scale systems. However, for broad commercial applications, the amount of training data required is more than that can be collected in a short time period. Therefore, a tradeoff exists between limiting the amount of data needed for model training and improvement in model performance by personalization.

We address this tradeoff by using clustering methods to separate different trust dynamics across the sample population. We then develop customized trust models for each cluster of the population that account for broad individual differences in trust dynamics but allows model development with limited data. Specifically, we consider an interaction between a driver and a Society of Automotive Engineers (SAE) Level 2 driving automation and collect self-reports of trust throughout the interaction. We identify groups of users with critical differences in their trust dynamics using clustering based on trust evolution features. We demonstrate that the customized models based on these clusters significantly outperform the general model in predicting human trust as well as their take-over behavior. Additionally, although demographic factors have significant contributions to individual differences, we show that clustering based on trust dynamics-based features is more effective than simple demographic factor-based clustering for trust behavior prediction. Finally, we support the existence of the resulting clusters with literature from behavioral psychology. In summary, the contributions of this work are:

- 1) a framework to cluster humans based on their trust behavior dynamics;
- 2) identification of the “confident” and “skeptical” groups of users based on their trust in automation that is grounded in literature; and
- 3) improvement in prediction performance of human trust and take-over behavior with limited data available using the clustering-based customized models.

¹ Jundi Liu is with the Industrial & Systems Engineering Department at the University of Washington, Seattle, WA 98105. This work was conducted during his internship at Honda Research Institute. jundiliu@uw.edu

² Kumar Akash, Teruhisa Misu, and Xingwei Wu are with Honda Research Institute USA, Inc., San Jose, CA 95134, USA. {kakash,tmisu,xingwei.wu}@honda-ri.com

To the best of our knowledge, this is the first study that clusters users based on their trust dynamics, leading to an improved customized model for real-time trust prediction.

II. RELATED WORK

With the fast emerging of vehicle automation technologies, there is skepticism rising in public concerns [17]. In a 2013 survey, 66% of U.S. respondents indicated they were “scared” by the concept of automated driving, and more than half of respondents are skeptical of the reliability of the technology [18]. The results show significant disagreements, that some people are more confident about the future of automated systems and others are still skeptical, have arisen in the attitudes toward the automated systems. From the human factors perspective, the vehicle automation needs to identify drivers’ psychological characteristics and cognitive processes because those factors are reported to influence how drivers use these technologies [19]. Trust has emerged as a relevant focus in research since it provides a solid foundation for describing the relationship between humans and automation [20], [17].

A. Trust understanding and modeling

Trust is shown to play a crucial role in understanding the acceptance of innovative technology [21]. In fact, [12] reported that trust determines the use or rejection of automation and willingness to rely on automation in certain situations. Researchers have pointed out important aspects of trust in human-machine interaction. Muir et al. [13] showed that an individual’s mental model has a strong relationship with the way he/she trust in the system. The research also emphasized that trust changes through experiences in association with the change of his/her mental model of the system.

A well-accepted definition of trust in automation, proposed by Lee and See [22], is “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability”. They propose that the dynamic process rules trust and introduces context as a significant factor in trust development. Based on the definition of trust in [22], Hoff and Bashir [12] introduce three layers of trust: dispositional, situational, and learned trust. The dispositional trust is conceptualized based on early trust-related experiences that are typically affected by an individual’s demographics. Situational trust is context-dependent and is affected by situational information. The learned trust evolves with the experience with the system and differs by individual’s mental model. This work is widely referenced in the later studies since they consider the trust evolution as a dynamic process [23], [24], [25].

B. Trust estimation and individual differences

Most studies adopt trust-related questionnaires to obtain users’ trust levels. Lee and Kolodge [26] applied a topic model-based clustering method to comments about consumer attitudes towards vehicle automation. However, the questionnaires have a delayed effect and ignore the trust dynamics. While the factors discovered in the study provide guidance

on feature extraction of trust dynamics, we aim at capturing individual differences in trust dynamics to achieve real-time prediction of user trust for trust calibration.

Morra et al. [27] proposed using a combination of questionnaire and galvanic skin response signal to quantify trust. The experiment was carried out using virtual reality as a human-machine interface to convey situational information, which is shown to improve trust in vehicle automation. Akash et al. [28] proposed a customized set of psychophysiological features for each individual to build a classifier-based trust-sensor model, which has investigated the trust estimation in real-time. Nevertheless, the psychophysiological measurements are intrusive and impractical in real-world implementation. Several researchers have developed a variety of quantitative human trust models. These include regression models [29], [30], time-series models [31], [22], [32], and Markov models [33], [34]. Recent work has demonstrated the use of a partially observable Markov decision process (POMDP) to model human trust dynamics to improve human-robot performance [35]. Researchers have also used a state-space model to capture human trust dynamics while interacting with a Level 3 driving automation based on automation performance, drivers’ gaze, and drivers’ non-driving related task performance [15].

With the advancement of new methodologies, many new approaches have been proposed in recent years. Although there are studies that consider cultural differences that affect trust in automated systems [36], previous studies on quantitative trust modeling often disregard the “dispositional” aspect. In this work, we demonstrate that individual differences in trust behavior (i.e., dispositional trust) can be captured by clustering the participants based on their trust dynamics (i.e., how they gain/lose the trust) to improve the performance of quantitative trust models.

III. ONLINE STUDY DESIGN

To model and cluster the dynamics of human trust, we collected human subject data using an online study where the participants interact with a Level 2 driving automation. The study used a simulated autonomous driving recording that was prerecorded using a physical driving simulator. The study was deployed on Amazon Mechanical Turk [37], and the participants accessed the study online using their personal computers. During the study, the autonomous car drove through a series of ten intersections in an urban environment. The participants could press the spacebar key on their keyboard to indicate their intent to take over if they did not feel safe with the driving. Along with their take-over behavior, participants were also asked to provide self-reports of their trust as well as the reliability of the automation after each intersection during the study. Additionally, they completed a 12-question 7-point Likert scale pre-study and post-study trust questionnaire adapted from [38].

Two within-subject factors were varied for the ten intersections: automation reliability and pedestrian presence. Automation reliability was defined only in terms of car’s stopping behavior at an intersection for consistency. It had

two levels: low reliability, where the car aggressively decelerates very close to the stopline (deceleration starting at < 25 meters), and high reliability, where the car smoothly decelerates toward the stopline (deceleration starting at > 60 meters). Note that the reliability of the driving automation can be varied by other factors as well. Pedestrian presence also had two levels: either pedestrians are present or absent at the intersection. The presence of pedestrians can increase the perceived risk by the drivers. The risk was randomly varied across the ten intersections to avoid any ordering effects.

Additionally, three factors that can affect human trust dynamics were varied between participants: scene visibility, overall reliability, and automation transparency. First, weather condition is shown to impact the trust levels significantly [39]. Scene visibility was varied by changing the weather of the environment and had two levels: high visibility where the scene was sunny and low visibility where the scene was foggy with snow. Second, situational characteristics such as automation reliability can also substantially impact trust dynamics [40]. In particular, users gain more trust in systems that are reliable. Overall reliability was designed to be affected by scene visibility such that the reliability of the automation reduces in low visibility scenes (as expected in real scenarios due to degraded scene perception). Specifically, the overall reliability had three levels: 100% reliable, where none of the intersections were of low reliability and only occurred during high visibility scenes; 80% reliable, where two intersections were of low reliability and occurred during both low and high visibility scenes; and 60% reliable, where four intersections were of low reliability and only occurred during low visibility scenes. The low reliability intersections were randomly chosen across the ten intersections. Third, studies have shown automation transparency also has a positive correlation with trust levels. With a higher level of transparency, the users have more access to the situational information leading to an increase in their trust [41]. Studies have shown changing the level of automation transparency can calibrate trust to the appropriate levels [35], [42]. In our study, during the high automation transparency, augmented reality (AR) cues are shown to the participant representing the driving automation’s perception of the scene. The AR cues can provide vehicle speed information, navigation information, and object detection and prediction that are present in the scene. In low transparency, the object detection and prediction are not presented. Fig. 1 shows an example screenshot of the actual study scenario. Tab. I shows the resulting eight drive types and their corresponding characteristics. Tab. II shows the randomized distribution of reliability and risk in each drive type. Refer to the supplementary video for further demonstrations.

Participants: Two hundred thirty nine participants (121 males, 113 females, and 5 unknown) with ages between 19 and 77 years (mean: 39 years) from the United States participated in and completed the study online. They were recruited using Amazon Mechanical Turk, with the criteria that they must live in the US and have completed more

TABLE I

EIGHT DRIVE TYPES IN THE ONLINE STUDY. OVERALL RELIABILITY IS THE PERCENTAGE OF HIGH RELIABILITY OPERATIONS.

Drv. Type	Overall Reliability	Visibility	Transparency
A	100%	High	High
B	80%	High	High
C	80%	Low	High
D	60%	Low	High
E	100%	High	Low
F	80%	High	Low
G	80%	Low	Low
H	60%	Low	Low



Fig. 1. An example screenshot of the study scenario of low visibility and high automation transparency. The bounding boxes highlight the signs, cars and other moving objects in sight.

than 1000 tasks with at least a 95% approval rate. The compensation was \$2.25 for their participation, and each participant electronically provided their consent. The Institutional Review Board at Purdue University approved the study. Each participant completed a randomly selected drive type from Tab. I. Before the participants began the trial, they were given brief instructions about the study, and they completed a tutorial consisting of four intersections that helped familiarize them with the study interface. To ensure a uniform notion of trust across participants, they were explicitly informed about the definition adapted from [22] as follows:

“Trust is defined as your attitude that the self-driving car will help you achieve your goal of driving safely in a situation characterized by uncertainty and vulnerability.”

Since the participants were not monitored during the study, we asked the participants to complete the study in fullscreen mode to avoid distractions. To avoid non-complying participants, we tracked the key-presses on the keyboard during the trial and removed the participants from the dataset who were suspected of exiting the fullscreen mode during the study. Furthermore, we removed the participants from the dataset who had missing survey data. As a result, 40 participants’ data were removed from the dataset.

To summarize the online study design, we collected real-time trust levels of the users along with their take-over behavior while interacting with a Level 2 driving automation in a simulated environment. The study design considers three drive-level factors that can significantly impact the drivers’ trust dynamics. Moreover, two event-level factors are considered within each drive type. The collected data will be used to identify the key characteristics of users trust dynamic, which will be the basis of clustering users based on their trust behavior.

TABLE II

EVENT CONFIGURATION IN EACH INTERSECTION. THE CROSSES DENOTE THE LOW RELIABILITY OPERATION, AND THE PS DENOTE THE PRESENCE OF PEDESTRIANS IN THE SPECIFIC INTERSECTION.

Drv. Type	1	2	3	4	5	6	7	8	9	10
A		P			P		P	P	P	
B		P		P	X P	P		X	P	
C		P			X P	P	P	P	X	
D			X	X P	X P		P		P	X P
E					P		P	P	P	P
F	P		P	P			P	P	X	X
G	P					X P	P	P		X P
H	P	P		X	P	X	P	X	X P	

IV. METHODOLOGY

Using the intersection-by-intersection trust measurements obtained from the participants, we analyze the trust dynamics of each individual. We observe significant variations across individual trust behavior. Specifically, some participants start with lower initial trust levels as they may be skeptical about the automation. As they interact with a consistently high reliability automation, their trust levels increase gradually. However, if they encounter a low reliability operation, their trust levels drop drastically. As an extreme case, some participants' trust levels remain low throughout the study and do not increase significantly even after experiencing high reliability operation. On the contrary, some participants start with high initial trust levels and maintain stable and high trust levels throughout the study.

A. Trust Evolution Decomposition

We visualized the collected trust data to gain insights into the trust dynamics. Since there are drive types that do not have the low reliability operations or the low reliability operations happen at the end of the drive, we remove the participants from those drives to ensure all the participants for analysis have encountered the similar events in the study. As a result, we removed drive type A, E, and F from the samples, which left with 138 participants for further analysis. Based on the observations from the participants' data, we noticed the trust dynamics had three general phases across all the participants. The first phase is the initial *trust-building* phase. In this phase, the participants start from their initial trust levels and gradually gain some trust as they experience high reliability operations. The initial trust levels depend on the personal characteristics of the participants. For example, people who have indicated prior good experience with advanced vehicle automation systems in the pre-study survey (scores are greater than 5 out of 7) are likely to have a higher initial trust level (average initial trust level is 88.6). The second phase is the *error-awareness* phase, which occurs when a participant encounters the low reliability operation of the automation. After observing that the automation is not perfect, participants typically lose their trust in automation drastically. Finally, the third phase is the *trust-repair* phase. This phase follows the error-awareness phase during the interaction, where the participants regain their trust in the automation as they experience consistent high reliability operations after low reliability ones. The trust

increase during the trust-repair phase is typically lower than the initial trust-building phase since the participants realize that the automation may not be perfect and is prone to errors.

For example, Fig. 2 shows the average trust dynamics for drive G. In this drive, there are two low reliability operations at intersections 6 and 10. We observe that participants typically start with a relatively high average initial trust level (~ 77). This is consistent with findings that recent widespread use of automation has led to humans trusting a system when they have no experience with it [43]. During the initial trust-building phase (i.e., during consistent high reliability operation till intersection 5), the average trust level gradually increases for most of the participants, as apparent from smaller confidence intervals. At intersection 5, the trust level has the smallest confidence interval with the highest average value for this drive. Then in the two error-awareness phases (intersection 6 and 10), we observe a significant decrease in the trust levels and wider confidence intervals compared to the previous intersection, respectively. In this phase, participants notice the low reliability operations, which significantly reduces their trust level. Moreover, the first low reliability operation results in a much more significant effect in the decrease of trust level. Finally, the trust-repair phase (intersection 7 to intersection 9) shows the participants gradually regaining their trust after the low reliability operation if no further low reliability operation occurs. However, the rate at which the trust increases during the trust-repair phase is not as large as during the initial trust-building phase. Similar trends are also observed in other drive types.

In summary, we observed three main phases of the trust dynamics during the interaction with imperfect automation. However, the wide confidence intervals in Fig. 2 also suggest a considerable variation of the participants' trust levels across the three phases. In particular, the wide confidence intervals for the initial trust level show the participants have significant initial differences toward the automation. Although the confidence intervals converge during the trust-building phase, the confidence intervals are still wide during the error-awareness and trust-repair phases. These observations suggest that a general model may not be effective in capturing the individual differences in trust dynamics, and there is a need to develop customized models of trust.

B. Trust Behavior Clustering

To cluster the participants based on their trust dynamics, we first extract features characterizing the dynamics and behaviors of participants' trust throughout each interaction for each individual. Specifically, we extract the rate of change of trust to capture the trust dynamics for each of the three trust evolution phases. Furthermore, based on our preliminary analysis of trust, we observe that trust is strongly correlated to the presence of pedestrians at the intersection as well as participants' take-over response. Therefore, for the trust-building and trust-repair phase, we extract the average trust of participants at intersections where: 1) pedestrians are present, 2) pedestrians are absent, 3) the participants take-

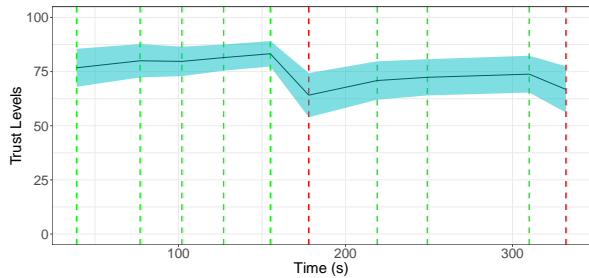


Fig. 2. Average trust dynamics for drive G. Each dashed vertical line denotes an intersection in the study. The green lines represent the high reliability operation, and the red lines represent the low reliability operation in that particular intersection. The black solid line denotes the average trust levels across all participants for drive G with the blue shaded area denoting the 95% confidence intervals.

TABLE III
LIST OF 12 EXTRACTED TRUST DYNAMICS FEATURES.

Phase	Feature
Trust-building	Initial trust level
	Rate of change of trust
	Average trust with pedestrian present
	Average trust with pedestrian absent
	Average trust with take-over
Error-awareness	Rate of change of trust
	Rate of change of trust
Trust-repair	Average trust with pedestrian present
	Average trust with pedestrian absent
	Average trust with take-over
	Average trust with no take-over
	Average trust with no take-over

over 4) the participants do not take over. We additionally consider the initial trust level as a feature as it can capture the individual differences we observed in our analysis. This results in 12 features in total as listed in Tab. III.

Considering the limited sample size of unique participants¹, we use simple Euclidean distance-based clustering method. To minimize the ‘curse of dimensionality’ [44] for the clustering algorithm, we use principal component analysis (PCA) to reduce the number of features used for clustering [45]. We apply PCA on the extracted features in Tab. III to reduce the dimension of data and provide insights on significant features that contribute most to the total variation. The explained variance ratios for the first three principal components (PC) are 44%, 17%, and 11%, respectively². Thus, the first 3 PCs explain about 72% of the total variance in the extracted features.

Finally, we use the K-means clustering method to find the groups of people with similar trust behaviors. K-means is one of the most widely used clustering methods, which iteratively computes each cluster’s centroid and updates the assignment of each sample. While converging to stable assignments, K-means finds the clusters which minimize within-cluster

¹Although our data is not small to model trust behavior as each participant contribute to multiple trust samples, the number of unique participants (138) is limited for data-based clustering.

²The variance ratios for the 4th and the subsequent components are 9%, 6%, etc.

variances. We chose the number of clusters as two since the silhouette analysis shows two clusters have the largest average silhouette coefficient (0.45) among all numbers of clusters varying from two to six and it ensures the best interpretability with significant statistical difference in all extracted features. With the identified clusters of participants, we will train customized models for each cluster to capture the individual differences of trust dynamics in each group. Furthermore, a close look into the clusters can provide insights into group-specific similarities.

V. RESULTS AND VALIDATION

We applied the K-means clustering algorithm on the first three PCs to generate two clusters. The resulting two clusters have 36 and 102 participants, respectively. Fig. 3 shows the boxplots that demonstrate the variations in four representative features for the identified two clusters. For the initial trust (Fig. 3(a)) and the average trust with pedestrian absence during trust-building (Fig. 3(b)), the two clusters show a statistically significant difference based on two sample t-test with a significant level of 0.05. Specifically, the cluster shown in orange has relatively lower initial trust than that shown in blue. Furthermore, the orange cluster also has a lower average trust with pedestrian absence during trust-building than the blue cluster. This shows that the participants comprising the orange cluster are more skeptical than those comprising the blue cluster. They tend to have a low initial trust toward the automation as well as a low trust level even during a low risk (due to the absence of pedestrians) and high reliability operations. Therefore, we name the orange cluster as the “skeptical” group due to their lack of trust. On the contrary, we call the blue cluster the “confident” group as they show high confidence on the driving automation. In the error-awareness phase, we observe from Fig. 3(c) that the “skeptical” group’s trust levels are more volatile compared to the “confident” group and thereby drop faster after they encounter a low reliability operation. Finally, during the trust-repair phase, we consider the average trust level with no take-over (Fig. 3(d)). This feature represents the trust level when the participants are comfortable with the automation system. We observe that the “skeptical” group has statistically lower trust levels than the “confident” group even when they do not take over. A two group t-test shows all 12 features are significantly different for the two clusters.

To better demonstrate the characteristics of the trust dynamics for the two groups, we compare the evolution of average trust for same drive type G as Fig. 2 in Fig. 4. The data for this drive type comprise of 22 “confident” group participants and 5 “skeptical” group participants. We see that the “skeptical” group start with a relatively lower initial trust and their trust level drops more quickly during a low reliability operation compared to the “confident” group.

While these clusters were identified using data-driven techniques, it is important to verify these behaviors with cognitive and behavioral psychology literature. Prior studies of both interpersonal human trust as well as human trust in automation note the existence of distinct trust behaviors

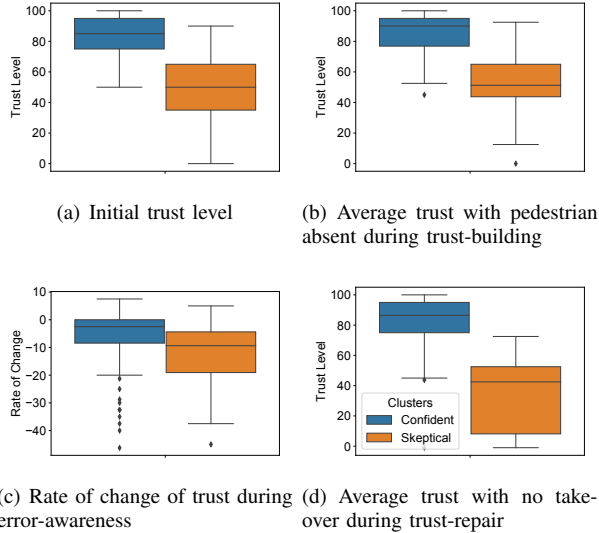


Fig. 3. Boxplots of four representative features of trust dynamics. The two clusters show significantly difference from each other in all four features.

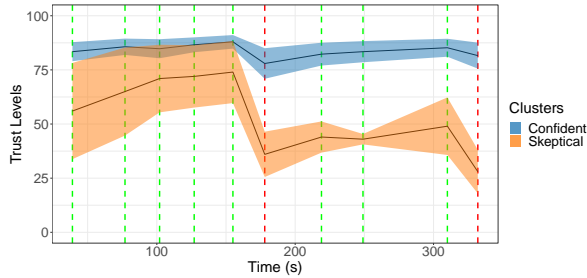


Fig. 4. Average trust level of the “skeptical” and the “confident” group participants at each intersection for drive G. The shaded area denotes the 95% confidence intervals of the trust level.

among humans. Studies using the Rotter Interpersonal Trust Scale [46] have found two groups of trust behaviors, namely, “high trusters” and “low trusters”. Furthermore, it was noted that characteristics of each individual varied in terms of willingness to trust a novel situation. Though the differences between humans’ trust in automation was not explicitly analyzed, the other metrics established the two groups, such as high trusters being more willing to trust experimenters [47], [48]. It is likely that the “confident” and “skeptical” partially represent the high and low trusters, respectively.

A. Customized Real-time Trust Prediction

We show the improvement in trust prediction performance using customized models as compared to general model. The general model is trained using data for all participants and therefore, ignores the individual trust variations in trust dynamics. The customized models are separately trained using data from each group of participants identified by clustering, respectively. The resulting customized models leverage the distinct trust behaviors of each group that significantly improve the models performance. We consider two structures of trust prediction models for this comparison.

1) *Linear Regression (LR) Model*: We first consider a simple linear regression model for predicting users’ trust. We

use the observations from the previous two intersections to predict the trust level at the current intersection. Specifically, we consider scene visibility, automation transparency, presence of pedestrians, automation reliability, participants trust level, and participants take-over behavior in the previous two intersections as an input to the model to predict the current trust level. Although this model may not be practical if self-reports of trust are unavailable, the models provides a simple baseline to validate the clustering performance.

2) *State Space (SS) Model with Kalman Filter*: A classical approach to model a dynamical system is by using a linear time-invariant state-space (SS) model. Linear SS model has been used to capture human trust dynamics while interacting with a Level 3 driving automation based on automation performance, drivers’ gaze, and drivers’ non-driving related task performance [15]. Considering trust as a continuous state of the system, they use Kalman filter to estimate trust during the interaction. Since our output of take-over intent is a binary variable, we adapt the linear state space model in [15] with a sigmoid output function that maps the state of trust to the output of take-over. Thereby, we consider scene visibility, automation transparency, pedestrian presence, and automation reliability as inputs of the SS model; the state is a continuous variable of trust; and the output is the take-over intent. The model parameters are estimated using linear mixed-effect model with participants as a random effect as described in [15]. Note that the self-reported trust is used as the measurement for the continuous trust state and is only used for model training. To evaluate the prediction performance of trust and take-over intent, we use an extended Kalman filter (EKF) that accommodates the nonlinear sigmoid output function to update the state estimate of trust after each output is observed, which is then used to predict the next take-over intent based on current inputs. Therefore, the EKF provides a real-time estimate of trust and take-over without the need for self-reports of trust during real-time prediction. The model is characterized by (1).

$$\begin{aligned} T_{k+1} &= \mathbf{A}T_k + \mathbf{B} [v_k \ t_k \ p_k \ f_k]^T \\ b_k &= \text{Sig}(\mathbf{C}T_k + \mathbf{C}_b) \end{aligned} \quad (1)$$

Here $\text{Sig}(x) = (1 + e^{-x})^{-1}$, T_k is the trust level at the k th event, v is the scene visibility, t is the automation transparency, p is the presence of pedestrians, f is the automation reliability, b is the take-over behavior of the participant, and \mathbf{A} , \mathbf{B} , \mathbf{C} , and \mathbf{C}_b are linear parameters.

Additionally, we compare the performance of our proposed trust dynamics-based clustering with demographic information based clustering. Specifically, we consider three baselines: 1) age with threshold as the mean age (40 years) of the participants sample; 2) gender (male or female); and 3) self-reported driving style (aggressive or conservative). We perform a 5-fold cross validation (CV) with uniform distribution of each drive type in the training and validation sets for both the model structures to obtain the validation performances. We calculate the validation mean squared errors (MSE) for trust prediction using both models and the

TABLE IV

MSE AND F1 SCORES FOR GENERAL MODEL AND CUSTOMIZED MODELS USING DIFFERENT CLUSTERING CRITERIA. LOWER MSE AND HIGHER F1 SCORE INDICATES BETTER MODEL PERFORMANCE.

Clustering Criteria	Cluster	Number of participants	MSE for trust		F1 scores for take-over
			LR model	SS model	SS model
-	General model	138	0.602	0.518	0.423
Trust dynamics (our method)	“Confident”	102	0.442	0.381	0.468
	“Skeptical”	36	0.384	0.289	0.650
Age	“At least 40”	53	0.587	0.586	0.520
	“Less than 40”	85	0.592	0.431	0.327
Gender	“Male”	65	0.594	0.488	0.456
	“Female”	71	0.526	0.549	0.396
Driving Style	“Aggressive”	61	0.543	0.571	0.347
	“Conservative”	77	0.526	0.493	0.460

TABLE V

PERCENTAGE INCREASE IN THE PREDICTION PERFORMANCE USING THE CUSTOMIZED MODEL AS COMPARED TO THE GENERAL MODEL

Clustering Criteria	MSE for trust		F1 score for take-over
	LR model	SS model	SS model
Trust dynamics	29.1%	31.1%	21.7%
Age	2.0%	5.2%	-5.2%
Gender	7.1%	-0.3%	0.5%
Driving style	11.3%	-1.7%	-3.1%

F1 scores for take-over intent prediction using the SS model for the general model as well as for the clusters using each of the clustering criteria. The result is shown in Tab. IV.

We see that for both the LR and SS models, the customized models based on the trust dynamics-based clustering performs better than the general model. That is, the trust dynamics-based customized models have lower MSE for trust as well as higher F1 score for take-over intent prediction. Therefore, the customized models successfully improves the trust prediction by considering the individual differences across the population. However, we do not see such significant improvement for age-based, gender-based, or driving style-based customized models. To further quantify the improvement in the prediction performance using the customized model, we calculate the percentage increase in the average metrics (MSE and F1 score) for each clustering criteria as compared to the general model. The average metrics for each clustering criteria is calculated as the mean of the metrics across the clusters weighted by the number of participants in each cluster. The resulting improvements in the prediction performance is shown in Tab. V. We observe that the trust-dynamics based customized models not only improves the prediction performance, it significantly outperforms simple demographic factor based clustering. This shows that these demographic factors alone may not be strong contributors to the variations in human trust behavior.

In summary, we demonstrate that trust dynamics-based clustering can allow to develop improved trust model needed for trust calibration paradigms. In practice, a small interaction data from a user can be used to determine the cluster to which the user belongs to and accordingly utilize the pre-

trained customized models and policies for the given cluster. This allows ease of deployment in commercial settings without the need to retrain the models for personalization.

VI. CONCLUSION

We presented a trust dynamics-based clustering framework to identify and develop customized trust models based on dominant human trust behavior among a large population. We showed that such a framework can balance the tradeoff between a single general, or several personalized, models of human trust. We identified participants in the two clusters, namely “skeptical” and “confident” based on their trust behavior. We showed that customized models developed based on these clusters significantly outperforms a general *one-fit-all* model in predicting human trust and take-over behavior during interaction with a driving automation. Furthermore, trust dynamics-based clustering approach is better than age-, gender-, or driving style-based approach in developing such customized model. Finally, we showed that the clustered participants’ behaviors could be explained reasonably and may be coincident with established psychology of human trust. Future work could involve using these customized models for real-time trust calibration paradigms to improve human-automation interactions.

ACKNOWLEDGMENT

We sincerely acknowledge Jain Research Lab and REID Lab at Purdue University for human subject study design and data collection.

REFERENCES

- [1] D. V. McGehee, M. Brewer, C. Schwarz, B. W. Smith, *et al.*, “Review of automated vehicle technology: policy and implementation implications.” Iowa. Dept. of Transportation, Tech. Rep., 2016.
- [2] U. Z. A. Hamid, F. R. A. Zakuan, K. A. Zulkepli, M. Z. Azmi, H. Zamzuri, M. A. A. Rahman, and M. A. Zakaria, “Autonomous emergency braking system with potential field risk assessment for frontal collision mitigation,” in *2017 IEEE conference on systems, process and control (icspc)*. IEEE, 2017, pp. 71–76.
- [3] S. Elmalaki, H.-R. Tsai, and M. Srivastava, “Sentio: Driver-in-the-loop forward collision warning using multisample reinforcement learning,” in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 28–40.

- [4] M. Kuderer, S. Gulati, and W. Burgard, "Learning driving styles for autonomous vehicles from demonstration," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2641–2646.
- [5] D. Niu, J. Terken, and B. Eggen, "Anthropomorphizing information to enhance trust in autonomous vehicles," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 28, no. 6, pp. 352–359, 2018.
- [6] E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman, "A design methodology for trust cue calibration in cognitive agents," in *International conference on virtual, augmented and mixed reality*. Springer, 2014, pp. 251–262.
- [7] K. Drnec and J. S. Metcalfe, "Paradigm Development for Identifying and Validating Indicators of Trust in Automation in the Operational Environment of Human Automation Integration," in *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*, D. D. Schmorrow and C. M. Fidopiastis, Eds. Switzerland: Springer International Publishing, 2016, vol. 9744, pp. 157–167.
- [8] J. S. Metcalfe, A. R. Marathe, B. Haynes, V. J. Paul, G. M. Gremillion, K. Drnec, C. Atwater, J. R. Estep, J. R. Lukos, E. C. Carter, and W. D. Nothwang, "Building a framework to manage trust in automation," in *Micro- and Nanotechnology Sensors, Systems, and Applications IX*, vol. 10194, May 2017, p. 101941U.
- [9] K. Akash, N. Jain, and T. Misu, "Toward adaptive trust calibration for level 2 driving automation," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 538–547.
- [10] P. A. Ruijten, J. Terken, and S. N. Chandramouli, "Enhancing trust in autonomous vehicles through intelligent user interfaces that mimic human behavior," *Multimodal Technologies and Interaction*, vol. 2, no. 4, p. 62, 2018.
- [11] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with Trust for Human-Robot Collaboration," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*. Chicago, IL, USA: ACM Press, 2018, pp. 307–315.
- [12] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human factors*, vol. 57, no. 3, pp. 407–434, 2015.
- [13] B. M. Muir, "Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.
- [14] A. Xu and G. Dudek, "OPTIMO: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: ACM, 2015, pp. 221–228.
- [15] H. Azevedo-Sa, S. K. Jayaraman, C. T. Esterwood, X. J. Yang, L. P. Robert, and D. M. Tilbury, "Real-time estimation of drivers' trust in automated driving systems," *International Journal of Social Robotics*, pp. 1–17, 2020.
- [16] P. Wintersberger, A.-K. Frison, A. Rienner, and L. N. Boyle, "Towards a personalized trust model for highly automated driving," *Mensch und Computer 2016-Workshopband*, 2016.
- [17] M. Hengstler, E. Enkel, and S. Duelli, "Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices," *Technological Forecasting and Social Change*, vol. 105, pp. 105–120, 2016.
- [18] K. Sommer, "Continental mobility study 2011," *Continental AG*, pp. 19–22, 2013.
- [19] M. Dikmen and C. Burns, "Trust in autonomous vehicles: The case of tesla autopilot and summon," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 1093–1098.
- [20] K. Raats, V. Fors, and S. Pink, "Understanding trust in automated vehicles," in *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, 2019, pp. 352–358.
- [21] M. Lüders, T. W. Andreassen, S. Clatworthy, and T. Hillestad, "Innovating for trust," in *Innovating for Trust*. Edward Elgar Publishing, 2017.
- [22] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.
- [23] K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems," *Human factors*, vol. 58, no. 3, pp. 377–400, 2016.
- [24] M. R. Endsley, "From here to autonomy: lessons learned from human-automation research," *Human factors*, vol. 59, no. 1, pp. 5–27, 2017.
- [25] S. M. Casner, E. L. Hutchins, and D. Norman, "The challenges of partially automated driving," *Communications of the ACM*, vol. 59, no. 5, pp. 70–77, 2016.
- [26] J. D. Lee and K. Kolodge, "Exploring trust in self-driving vehicles through text analysis," *Human factors*, vol. 62, no. 2, pp. 260–277, 2020.
- [27] L. Morra, F. Lamberti, F. G. Praticó, S. La Rosa, and P. Montuschi, "Building trust in autonomous vehicles: role of virtual reality driving simulators in hmi design," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 9438–9450, 2019.
- [28] K. Akash, W.-L. Hu, N. Jain, and T. Reid, "A classification model for sensing human trust in machines using eeg and gsr," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 4, pp. 1–20, 2018.
- [29] P. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," vol. 58, no. 6, pp. 719–735, June 2003.
- [30] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.
- [31] N. Moray, T. Inagaki, and M. Itoh, "Adaptive automation, trust, and self-confidence in fault management of time-critical tasks," *Journal of Experimental Psychology: Applied*, vol. 6, no. 1, pp. 44–58, 2000.
- [32] W. Hu, K. Akash, T. Reid, and N. Jain, "Computational Modeling of the Dynamics of Human Trust During Human-Machine Interactions," *IEEE Transactions on Human-Machine Systems*, pp. 1–13, 2018.
- [33] M. E. G. Moe, M. Tavakolifard, and S. J. Knapskog, "Learning trust in dynamic multiagent environments using HMMs," in *Proceedings of the 13th Nordic Workshop on Secure IT Systems (NordSec 2008)*, 2008.
- [34] E. ElSalamouny, V. Sassone, and M. Nielsen, "HMM-based trust model," in *International Workshop on Formal Aspects in Security and Trust*. Springer, Berlin, Heidelberg, 2009, pp. 21–35.
- [35] K. Akash, G. McMahon, T. Reid, and N. Jain, "Human trust-based feedback control: Dynamically varying automation transparency to optimize human-machine interactions," *IEEE Control Systems Magazine*, vol. 40, no. 6, pp. 98–116, 2020.
- [36] S. Hergeth, L. Lorenz, J. F. Krems, and L. Toenert, "Effects of take-over requests and cultural background on automation trust in highly automated driving," in *Proceedings of the Eighth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*. Salt Lake City, Utah, USA: University of Iowa, 2015, pp. 331–337.
- [37] Amazon, "Amazon Mechanical Turk," *Amazon Mechanical Turk - Welcome*, 2005.
- [38] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.
- [39] S. Sheng, E. Pakdamanian, K. Han, B. Kim, P. Tiwari, I. Kim, and L. Feng, "A case study of trust on autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 4368–4373.
- [40] B. E. Noah and B. N. Walker, "Trust calibration through reliability displays in automated vehicles," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 361–362.
- [41] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci, "Intelligent agent transparency in human-agent teaming for multi-uxv management," *Human factors*, vol. 58, no. 3, pp. 401–415, 2016.
- [42] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah, "Evaluating effects of user experience and system transparency on trust in automation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2017, pp. 408–416.
- [43] S. M. Merritt and D. R. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Human Factors*, vol. 50, no. 2, pp. 194–210, 2008.
- [44] I. Assent, "Clustering high dimensional data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 340–350, 2012.
- [45] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

- [46] J. B. Rotter, "A new scale for the measurement of interpersonal trust," *Journal of Personality*, vol. 35, no. 4, pp. 651–665, 1967.
- [47] —, "Generalized expectancies for interpersonal trust," *American Psychologist*, vol. 26, no. 5, pp. 443–452, 1971.
- [48] —, "Interpersonal trust, trustworthiness, and gullibility," *American Psychologist*, vol. 35, no. 1, pp. 1–7, 1980.