Dynamic Modeling of Trust in Human-Machine Interactions

Kumar Akash, Wan-Lin Hu, Tahira Reid, and Neera Jain

Abstract— **In an increasingly automated world, trust between humans and autonomous systems is critical for successful integration of these systems into our daily lives. In particular, for autonomous systems to work cooperatively with humans, they must be able to sense and respond to the trust of the human. This inherently requires a control-oriented model of dynamic human trust behavior. In this paper, we describe a gray-box modeling approach for a linear third-order model that captures the dynamic variations of human trust in an obstacle detection sensor. The model is parameterized based on data collected from 581 human subjects, and the goodness of fit is approximately 80% for a general population. We also discuss the effect of demographics, such as national culture and gender, on trust behavior by re-parameterizing our model for subpopulations of data. These demographic-based models can be used to help autonomous systems further predict variations in human trust dynamics.**

I. Introduction

Motivation and Problem Definition: The prevalence of autonomous systems facilitates process efficiency in both complex systems (*e.g.,* aircraft) and devices in daily life (*e.g.,* automated teller machines). Among various strategies studied to optimize automated processes, improving collaboration between humans and these systems is of great importance. This is because the benefits of automation may be lost when a human overrides an automated decision due to a fundamental lack of trust in a machine [1], [2]. Moreover, accidents may occur due to mistrust [3]. We aim to design autonomous systems that sense and respond to the trust levels of the humans they interact with, thereby resulting in a more productive relationship between the human and autonomous system. In order to achieve this goal, we need a model of dynamic human trust behavior that could be integrated into a feedback control system for improving the system's response to human trust.

The trust between humans is not necessarily the same as the trust of humans in autonomous systems. Scientists have adapted elements characterizing trust in social psychology and have investigated trust in human-machine interactions (HMI) and human-computer interactions (HCI) [1], [4], [5]. Most research has been focused on examining the significance of a specific factor (like gender) between different trust behaviors [6], [7]. However, defining statistically significant factors itself is insufficient for incorporating the factors into a control system. Furthermore, to develop a control system for general HMI usage, there is a need to model human trust based on human subject study data. The model should elicit the trust dynamics in all classes of human-machine collaborations.

Gaps in Literature: In order to derive a dynamic model of human trust behavior that is suitable for HMI and HCI contexts, an appropriate experimental design, modeling, and verification is necessary. There is no experimentally verified model for describing the comprehensive dynamics of human trust level in HMI contexts. Existing trust models are either nonlinear or do not capture the human behavior that is not based on rationale [8]. They also ignore the influence of the *cumulative effect* of past interactions on the present trust level. Finally, humans have different behaviors that are influenced by their surroundings and experiences. These are in turn strongly influenced by demographics of the particular human. With increasing globalization, autonomous systems will be deployed in different societies. Therefore, the ability to model human behavior for different demographics is a necessity for autonomous systems. There does not exist a generalized model structure in the present literature that can be adapted to these variations in human trust behavior.

Contribution: In this paper we propose an experiment which captures the *dynamic changes* in human trust, specifically in a HMI context. We establish a generalized linear time-invariant (LTI) model structure for human trust that is grounded in existing psychology literature. The simplicity of the proposed model makes it suitable for integration with feedback control systems. This will enable autonomous systems to respond to human trust variations accordingly. Supported by a large set of human behavioral data, we use gray-box system identification techniques to estimate the parameters of the trust model. We further systematically incorporate individual demographic factors by re-parameterizing our generalized model based on national culture and gender.

Outline: This paper is organized as follows. Section II provides background on trust modeling. The experimental procedure and behavioral response acquisition from human subjects are described in Section III. The generalized model description and methodology for model parameter estimation are presented in Section IV. Results and discussions are presented in Section V, followed by concluding statements in Section VI.

II. Background

Trust in HMI and HCI contexts has been studied extensively [9], [10]. Broadly speaking, trust in autonomous systems depends on a number of human, environmental,

*This material is based upon work supported by the National Science Foundation under Award No. 1548616. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

School of Mechanical Engineering, Purdue University, West Lafayette, IN 47906 USA kakash@purdue.edu, hu188@purdue.edu, tahira@purdue.edu, neerajain@purdue.edu

and robot/machine characteristics [11]. Trust itself can be classified into three categories: dispositional, situational, and learned [10]. Dispositional trust is based on characteristics of the human such as culture, gender, age, and personality. Situational trust consists of those factors that are external to the human (*e.g.,* task difficulty) and those that are internal to the human (*e.g.,* domain knowledge) [10]. Learned trust is based on the accumulation of experiences with autonomous systems and influences the initial mindset of the human. During a new *dynamic interaction* with an autonomous system, new experiences will be learned that will either reinforce the prior experiences or update them with new ones. In this paper, we will establish a human trust model that can capture learned and dispositional trust characteristics.

A. Studies on Human Trust Modeling

Researchers have modeled human trust based on the human's past experiences, which forms an integral part of learned trust. Jonker and Treur introduced two types of trust dynamics: trust evolution functions and trust update functions [5]. Trust evolution functions map a sequence of trust related events (experiences) to a current trust level, while trust update functions generate the next trust representation based on a current trust level and a current experience. In order to verify the proposed trust dynamics, Jonker *et al.* conducted follow-up human subject experiments. The results suggested that the temporal dynamics of trust depend on positive or negative experiences. However, limited by the number of trials (10 trials), these studies only induced a single transition in trust level, making the comprehensive understanding of trust dynamics inconclusive. Additional limitations include the fact that this experiment only considered learned trust factors and the influence of a person's trust in an organization or object.

Other researchers have modeled human trust in the context of HMI. Lee and Moray used a simulated semi-automatic juice plant environment to characterize the users' changes in trust level [4]. The authors observed that performance, trust, and faults affected the level of trust, and used an ARMAV (Auto Regressive Moving Average Vector) analysis to model the input-output relationship of the system. Their follow-up study further showed automation is used when trust exceeds self-confidence [12]. These pioneering efforts demonstrated the effect of situational and learned trust on the interactions between humans and autonomous systems; however, the generality of their model was limited due to a small sample size (*i.e.,* four to five participants in each group), and the standard deviation of the data used in their regression model was considerably large.

More recently, researchers have incorporated elements that are not based on rationale in the human trust model. Hoogendoorn *et al.* introduced 'bias' into their model to account for this [8]. They formulated models with biased experience and/or trust and then validated these models via a geographical areas classification task. Their result suggested that biased model is capable of estimating trust more accurately than models without an explicit bias incorporated. However, their model was nonlinear in trust and experience, making it less desirable for use in control design.

B. Factors that Influence Trust

Apart from experiences, human trust behavior is also influenced by demographics including culture and gender. National culture consists of the values, behaviors, and customs that exist within the population of a country. One of the most comprehensive studies of national culture is Hofstede's six cultural dimensions which include Uncertainty Avoidance Index (UAI) [13], [14]. National culture affects the cognitive processes of building trust, so people from different cultures are likely to use different mechanisms to form trust [15] and show particular trust behavioral intentions [16]. Although national culture has a significant effect on human trust behaviors, little research examines its effects on trust in automation. Huerta *et al.* found that Americans were less likely to trust autonomous (decision-aid) systems than Mexicans in a fraud investigation scenario [17]. In one study, Americans' tendency to trust less in automated systems was observed in their attitudes towards "auto-pilots" while compared with Indians [18]. Gender differences in trust have been studied, particularly in economic contexts [6], [19], [20]. Although the gender effect on trust in automation has not been studied as thoroughly as it has been in economic studies, some studies showed gender differences in HCI and human-robot interaction (HRI) contexts [7], [21], [22].

In summary, published literature lacks comprehensive modeling of human trust dynamics. Existing experiments do not induce continuous and multiple transitions in trust level over time. Moreover, the influences of demographic factors on trust behavior have only been discussed in literature, but not modeled. The effects of different trust factors on dynamic behavior of human trust remain unexplored. We will address these key gaps in the following sections.

III. Human Subject Study

The focus of our experimental design is to capture how autonomous system performance as well as the humans' demographic background influence trust dynamics in HMI contexts.

A. Stimuli and Procedures

In this study, participants interacted with a computer-based simulation in which they were told that they would be driving a car equipped with an image based *obstacle detection sensor*. The sensor would detect obstacles on the road in front of the car, and the participant would need to repeatedly evaluate the algorithm report and choose to either trust or distrust the report based on their experience with the algorithm. The instructions specifically informed the participant that the image processing algorithm used in the sensor was in beta testing.

There were two equally probable stimuli: 'obstacle detected' and 'clear road'. Participants could choose 'trust' or 'distrust' after which they received feedback of 'correct' or 'incorrect'. The trials were divided into two classes:

reliable and faulty. In reliable trials, the algorithm accurately identified the road condition, which was in fact the stimuli. For the participant, this meant that selecting 'trust' would be marked as 'correct' and selecting 'distrust' would be marked as 'incorrect'. For the faulty trials, there was a 50% probability that the algorithm incorrectly identified the road condition. Fig. 1 shows the sequence of events in a single trial.

Each participant completed four initial practice trials followed by 100 trials. The trials were divided into three phases, called 'databases' in the study as shown in Fig. 2. Before the start of each database, participants took a break of 30 seconds. In database 3, the accuracy of the algorithm was switched between reliable and faculty according to a pseudo-random binary sequence (PRBS) in order to excite all possible dynamics of the participant's trust responses. The order of reliable and faulty trials was counterbalanced on a between group basis in groups 1 and 2 (see Fig. 2).

B. Participants

Five hundred eighty-one participants (340 males, 235 females, and 6 unknown) recruited using Amazon Mechanical Turk [23], ranging in age from 20–73 (mean 35.32 and standard deviation 10.84) participated in our study. The compensation was $\$0.50$ for their participation, and each participant electronically provided their consent. The Institutional Review Board at Purdue University approved the study. We collected participants' demographic information via a post-study survey which included information about their gender along with the country in which they grew up. The latter is defined as national culture in this study.

IV. Modeling

In this section we discuss how the discrete responses of human participants were processed. We also present the generalized trust model that we will later parameterize using the human participant data.

A. Data Processing

We identified outliers and removed them from the data set according to the interquartile range (IQR) rule (the $1.5\times$IQR rule) [24]. We calculated the first quartile (Q_1) and the third quartile (Q_3) for three categories of data: the number of trust responses, distrust responses, and no responses. As a result, 63 outliers were removed from the dataset out of a total of 581 participants.

We calculated the *probability of trust response* for each trial across all subjects in groups 1 and 2. The probabilities varied from approximately 0.5 to 1 with 0.5 representing low trust and 1 representing high trust. These probabilities of trust, that varied with the evolution of the scenario, could be interpreted as the level of trust for the sample population and hereafter will be referred to as *trust level* $T(n)$, where $n \in [1, 100]$ is the trial order. Similarly, We calculated the *probability of reliable performance* for each trial across all subjects in groups 1 and 2. The probabilities varied from approximately 0.5 to 1 with 0.5 representing negative experience and 1 representing positive experience and hereafter will be referred to as *experience* $E(n)$. Thus we obtain the dynamic variation of trust level $T(n)$ with experience $E(n)$ for all participants. In order to reduce noise from the dynamically varying signal, $T(n)$, we used the Savitzky-Golay filter with order 3 and window of size 5 [25]. The variation of trust level and experience as a function of trial number is shown in Fig. 3.

TABLE I

P-VALUES OBTAINED FROM A PAIRED T-TEST BETWEEN THE DATASET OF ALL PARTICIPANTS AND EACH DEMOGRAPHIC BIN

Dataset	All	US	India	Female	Male
All	-	0.0000*	0.0000*	0.0001*	0.0063*
US		-	0.0000*	0.3178	0.0000*
India			-	0.0000*	0.0000*
Female				-	0.0007*
Male					-

* $p < 0.05$; pairs are significantly different

We divided the responses of the participants into bins based on their demographics: two bins based on national culture (United States (US) versus India) and two bins based on their gender (male versus female). In order to determine the differences between the dataset of all participants and each of the demographic-specific datasets, or bins, we conducted paired t-tests between all five datasets. The results (see Table I) show that the dataset consisting of all participants is significantly different from each of the individual demographic bins. This indicates the necessity of tuning parameters on the basis of demographics. Moreover, the t-test results indicate a potential coupling effect between US and female.

In order to decouple the effect of one demographic on another, *e.g.,* effect of country on gender, the number of participants from both sections of the former demographic were equalized for the later demographic. This was done by randomly choosing a smaller set from the demographic with a larger sample population. For example, for the female bin, the number of Indian females and US females were equalized by selecting a random smaller set of US females as the number of US females were larger than number of Indian females in the sample population.

B. Trust Model Description

Most of the previously developed human trust models observed trust to be directly related to experience. A well-known model presented in [5] described the change in trust to be proportional to the difference of experience and trust. However, in addition to experience, we recognized the significance of the cumulative perception of trust and the human's expectations of the autonomous system in our pilot studies. Therefore, we adapted Jonker's model and introduced two additional states—Cumulative Trust (C_T) and Expectation Bias (B_X)—to accommodate the bias in human behavior due to human's perception of past trust and their expectations as shown in (1). The proposed model is a three–state model as compared to a single–state model in Jonker et

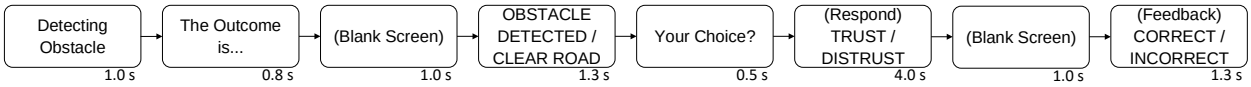


Fig. 1. Sequence of events in a single trial. The time length marked on the bottom right corner indicates the time interval that the information appeared on the screen.

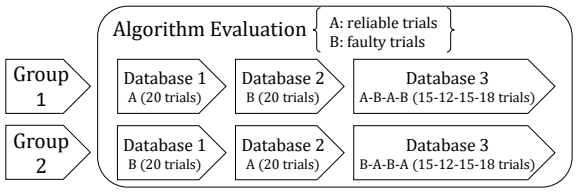
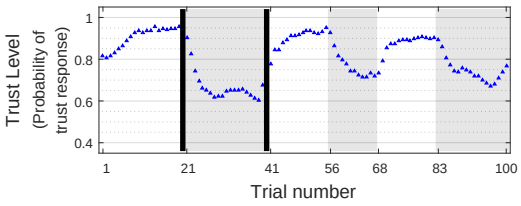
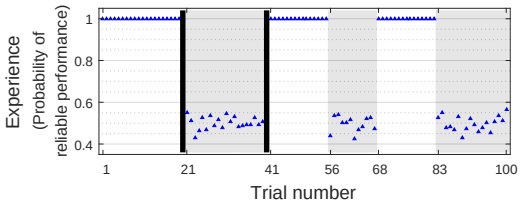


Fig. 2. Participants were randomly assigned to one of the two groups. The ordering of the three experimental sections (databases), composed of reliable and faulty trials, were counterbalanced across groups.



(a) Variation of trust level as a function of trial number.



(b) Variation of experience as a function of trial number.

Fig. 3. The trust level and the experience for all participants. Faulty trials are highlighted in gray, and black lines mark the breaks. Participants showed trust in reliable trials and distrust in faulty trials.

al. All of the variables and parameters in the model belong to $[0, 1]$ with $E(n)$ as the input and $T(n+1)$ as the output of the model.

$$
\begin{aligned}
T(n+1) - T(n) &= \alpha_e[E(n) - T(n)] && \text{(1a)}\\
&\quad + \alpha_c[C_T(n) - T(n)] && \text{(1b)}\\
&\quad + \alpha_b[B_X(n) - T(n)] && \text{(1c)}\\
C_T(n+1) &= [1-\gamma]C_T(n) + \gamma T(n) && \text{(1d)}\\
B_X(n+1) &= B_X(n) && \text{(1e)}
\end{aligned}
$$

In the proposed model (1), change in trust $T(n+1) - T(n)$ linearly depends on three terms: $E(n) - T(n)$ (1a), $C_T(n) - T(n)$ (1b), and $B_X(n) - T(n)$ (1c), where each term is bounded between -1 and 1. The difference between experience $E(n)$ and present trust level $T(n)$, $E(n) - T(n)$ (1a), updates the predicted trust level $T(n+1)$, so that it approaches $E(n)$. If the present experience is less than present trust level, then $E(n) - T(n) < 0$, thereby decreasing the predicted trust level and vice-versa.

We defined cumulative trust C_T as an exponentially weighted moving average of past trust level as shown in (1d). Cumulative trust incorporates the learned trust in the model using a weighted history of past trust levels. A higher value of the parameter γ discounts older trust levels faster, and thus γ can be called the *trust discounting factor*. The difference between present cumulative trust $C_T(n)$ and present trust level $T(n)$, $C_T(n) - T(n)$ (1b), updates the predicted trust level $T(n+1)$, so that it approaches $C_T(n)$.

Expectation bias B_X accounts for a human's expectation of a particular interaction with an autonomous system. This is meant to be constant during an interaction, as shown in (1e), but could change between different interactions. The difference between expectation bias $B_X(n)$ and present trust level $T(n)$, $[B_X(n) - T(n)]$ (1c), accounts for the discrepancy between the expectation bias and the actual trust level of the human; it updates the predicted trust level $T(n+1)$ so that it approaches $B_X(n)$. The parameters α_e, α_c, and α_b determine the weights given to each of the difference terms in the rate of change of the trust level. We call α_e, α_c, and α_b the *experience rate factor*, *cumulative rate factor*, and *bias rate factor*, respectively, as they control the rate by which each individual difference affects the predicted trust level.

Since the model (1) is linear, it can be represented in state space form as shown in (2). The state space representation makes the proposed trust model amenable to design, synthesis, and implementation of simple control architectures.

$$
\begin{aligned}
x(n+1) &= \begin{bmatrix} (1-\alpha) & \alpha_c & \alpha_b \\ \gamma & (1-\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix} x(n) + \begin{bmatrix} \alpha_e \\ 0 \\ 0 \end{bmatrix} u(n)\\
y(n) &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x(n)
\end{aligned}
\qquad (2)
$$

where

$$
x = \begin{bmatrix} T \\ C_T \\ B_X \end{bmatrix}, \qquad u = E, \qquad \alpha = \alpha_e + \alpha_c + \alpha_b.
$$

C. Parameter Estimation

In order to identify the optimal parameter set, we implemented nonlinear least squares estimation using the function *nlgreyest* in the System Identification Toolbox (version 9.4) from MATLAB 2016a. We estimated parameters using 1) the data of all participants and 2) the data in each of the four demographic bins. Each dataset consisted of data from each of the three 'databases' in both group 1 (in which participants were initially faced with reliable trials) and group 2 (in which participants were initially faced with faulty trials). Parameter

estimation was conducted on multi-batch data by treating each 'database' as a batch because participants were given short breaks between databases during the experiment.

Finally, it is worth noting that the quality of any data-based parameter estimation is only as good as the data itself. In the context of human subject data, no number of samples can fully represent the human population. In order to calculate the possible error in parameter estimation caused by the variation in sample selection, we iterated the estimation 1000 times, with each iteration using a new randomly selected subset of data representing 90% of the total dataset for all participants and each demographic bin. The errors caused by the variation in sample selection for a 95% confidence interval were less than 2% for all of the parameters (Table II). Thus, the obtained parameters are robust to variations in the sample selection.

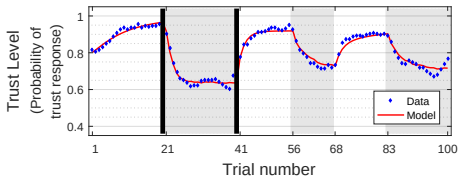
V. Results and Discussions

We verified our experimental design and the prediction model of human trust dynamics by fitting the model structure for a general population, which included all 581 valid participants in our experiment. Fig. 4 shows the experimentally obtained trust level and the predicted values using the trust model. The goodness of fit between the data and the model was calculated using the NRMSE (normalized root mean square error), and was 83.77% and 76.17% for group 1 and 2, respectively.

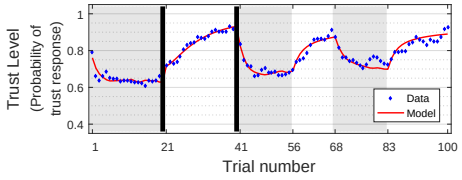
A. General Trust Behavior Observations

The study elicited the variation of trust level as expected: participants showed high trust level (*i.e.,* probability of trust response) in reliable trials and low trust level in faulty trials. This was achieved without training participants or providing them with specific information (*e.g.,* a game rule or background stories). The dynamics were modeled based on past behavioral responses and the experience of the human, and the prediction capability of the model was consistent irrespective of the initial condition of the system performance. In other words, the prediction capability of the model was consistent for both groups 1 and 2. This implies that the collaboration between the human and machine was the most significant factor in temporal variations in trust level. Therefore, the developed study is effective for modeling dynamic human trust behavior in HMI contexts.

The proposed study design induced trust dynamics by manipulating multiple transitions between positive and negative experiences. We observed that it took approximately eight to ten trials for participants to establish a new trust level (*i.e.,* approach steady state). The trust level still mildly increased or decreased near the steady state in both reliable and faulty trials. This finding was contrary to Jonker *et al.* [26] who asserted that "after a negative experience an agent will trust less or the same amount, but never more". Jonker's study was composed of only two sets of five trials, each with one transition in between, which we found to be less than the required number of trials to reach a steady state trust level.



(a) Group 1; Goodness of fit= 83.77%



(b) Group 2; Goodness of fit= 76.17%

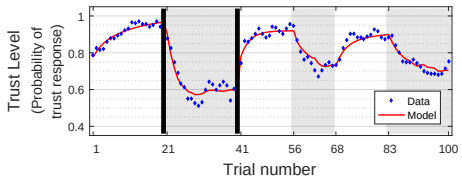
Fig. 4. Participants' trust level (blue dots) and the prediction (red curve) based on past behavioral responses and the experience of all participants. Faulty trials are highlighted in gray, and black lines mark the breaks between databases.

B. Effects of National Culture and Gender

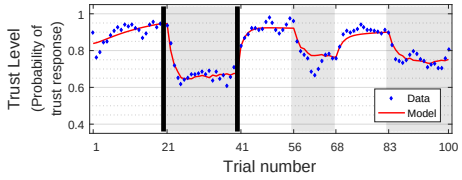
As expected based on the statistical analysis presented in Table I, different demographic groups show significantly different trust behavior from one another ($p = 0.00$ for US vs India and $p = 0.00$ for female vs male). We also observe that demographics such as national culture (Fig. 5) and gender (Fig. 6) had an effect on the participants' trust level towards the obstacle detection sensor in terms of rise time and steady state value.

Participants from the US trusted the autonomous system less when they could not determine whether the system was faulty or reliable in database 3 (see Fig. 5(a) and 5(c)). This is consistent with the findings that Americans trust autonomous systems less than Mexicans and Indians [17], [18]. Furthermore, the rate of change of trust was faster for Indian participants than US participants, and Indian participants also reached a higher trust level as compared to US participants. This agrees with the smaller Uncertainty Avoidance Index of Indian culture as compared to that of US culture (40 vs. 46) [27], which indicates that Indians are more tolerant of imperfection. Considering variations based on gender, males trusted more than females, especially when the system did not perform well (see Fig. 6(b) and 6(d)). This has also been discovered in several studies (*e.g.,* [19], [20]). On the other hand, females changed their trust level more rapidly than males.

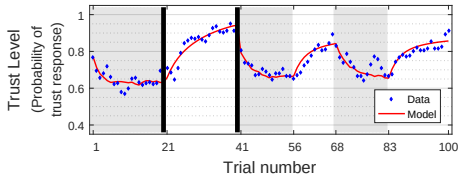
Based on the differences discussed above, the autonomous system would be able to interact with humans more appropriately based on customized models tuned to individual demographics. Table II shows the optimal parameter values for all participants and each demographic bin. The last two columns of Table II show the goodness of fit for each model with their corresponding group 1 and group 2 data. The goodness of fit for 'all participants' was more superior than that of individual demographic bins. This is due to the smaller sample size for individual demographics which resulted in more variations in probabilities.



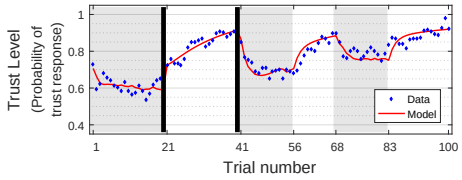
(a) US Group 1; Goodness of fit= 78.02%



(b) India Group 1; Goodness of fit= 68.99%

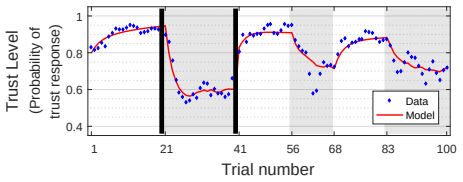


(c) US Group 2; Goodness of fit= 64.77%

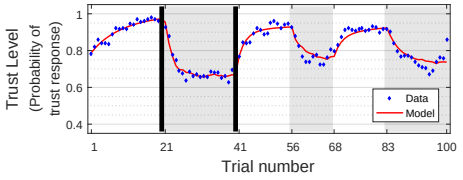


(d) India Group 2; Goodness of fit= 68.46%

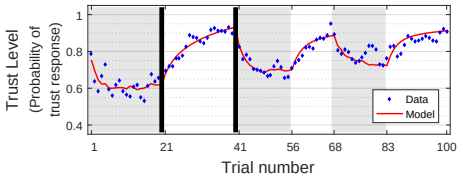
Fig. 5. Participants are grouped by the country where they grew up. Blue dots are the reported trust level while the red curve is the prediction from model. Faulty trials are highlighted in gray and black lines mark the breaks between databases.



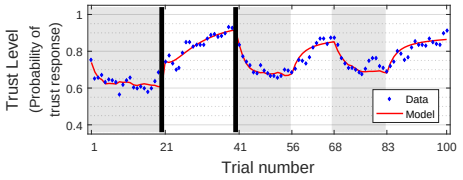
(a) Female Group 1; Goodness of fit= 69.98%



(b) Male Group 1; Goodness of fit= 73.95%



(c) Female Group 2; Goodness of fit= 65.80%



(d) Male Group 2; Goodness of fit= 68.48%

Fig. 6. Participants are grouped by their gender. Blue dots are the reported trust level while the red curve is the prediction from model. Faulty trials are highlighted in gray and black lines mark the breaks between databases.

Our modeling approach enables us to quantify this difference using two kinds of parameters—the rate factors and the discounting factor. Table II summarizes the identified parameters of the models. While comparing between countries, the net rate factor α ($\alpha_e + \alpha_c + \alpha_b$) for Indian participants is 27.8% larger than that of US participants. This implies that Indian participants' trust level increases or decreases faster than US participants' trust level after the system performance changes. Moreover, the trust discounting factor γ is 14.5% larger for US participants, indicating that US participants value their recent experience and information more as compared to Indian participants.

Similarly while comparing between genders, the net rate factor α ($\alpha_e + \alpha_c + \alpha_b$) for female participants is slightly larger (4.4%) than that of male participants, resulting in a slightly faster rate of change of trust for female participants. Additionally, the trust discounting factor γ is larger for male participants, indicating that male participants value their recent experience and trust more as compared to female participants. This finding partially agrees with Haselhuhn *et al.* who suggested that women are more likely to restore trust after a trust violation [28]. However, Haselhuhn *et al.* found that womens' trust decreases less than that of men as they prefer to maintain interpersonal relationships. This is contrary to our observations and is possibly due to their experimental context being different from ours. These results highlight that the dynamics of human trust behavior in HMI contexts is different from their interpersonal trust behavior, thus creating a need for human trust models in HMI contexts.

In summary, we derived a third-order linear model of human trust dynamics based on an experiment that elicited trust dynamics among more than 500 human subjects. Our generalized human trust model can provide an autonomous system with a mathematical characterization of a human's learned trust which is dynamically influenced by the system's performance. Moreover, by knowing the demographic information of the human, a system can further identify dispositional trust effects and facilitate the interaction accordingly. Although the number of input demographic factors is limited to one in the current model, more demographics can be incorporated by collecting more behavioral data for each demographic, *e.g.,* US males. This model can also be implemented with an online system identification algorithm in order to update the model in real-time for a given human.

TABLE II

OPTIMAL PARAMETER VALUES FOR ALL PARTICIPANTS AND EACH DEMOGRAPHIC BIN

Bin	Experience rate factor α_e	Cumulative rate factor α_c	Bias rate factor α_b	Trust discounting factor γ	Fit percentage Group 1	Fit percentage Group 2
All	0.1130 ± 0.0003	0.1019 ± 0.0009	0.1471 ± 0.0005	0.1465 ± 0.0009	83.80 ± 0.04	76.16 ± 0.06
US	0.1096 ± 0.0004	0.0716 ± 0.0005	0.1252 ± 0.0008	0.1201 ± 0.0009	77.90 ± 0.07	64.78 ± 0.13
India	0.1148 ± 0.0008	0.1273 ± 0.0028	0.1494 ± 0.0015	0.1049 ± 0.0013	68.86 ± 0.17	68.43 ± 0.12
Female	0.1185 ± 0.0007	0.1014 ± 0.0015	0.1348 ± 0.0013	0.1197 ± 0.0024	69.99 ± 0.13	65.89 ± 0.16
Male	0.1054 ± 0.0004	0.0858 ± 0.0007	0.1485 ± 0.0009	0.1343 ± 0.0015	73.94 ± 0.12	68.55 ± 0.08

VI. CONCLUSION

To attain synergistic interactions between humans and autonomous systems, it is necessary for autonomous systems to sense human trust level and respond accordingly. This requires autonomous systems to be designed using dynamic models of human trust that capture both learned and dispositional trust factors. In this paper, we described an experiment to elicit the dynamic change in human trust with respect to HMI contexts. We established a third-order linear trust model, grounded in existing psychology literature, which we then parameterized using human subject data collected from over 500 participants. In particular, we introduced two important states, cumulative trust and expectation bias, to more accurately capture human trust dynamics. The model elegantly captured the complex dynamics of human trust behavior and described differences in the trust behavior among different demographics. In particular, while maintaining a uniform model structure, we showed statistically significant differences in the model parameterization for different demographics.

REFERENCES

[1] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, vol. 27, no. 5, pp. 527 – 539, 1987.

[2] T. B. Sheridan and R. Parasuraman, "Human-automation interaction," *Reviews of Human Factors and Ergonomics*, vol. 1, no. 1, pp. 89–129, 2005.

[3] M. Richtel and C. Dougherty, "Google's driverless cars run into problem: Cars with drivers," *New York Times*, vol. 1, 2015.

[4] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[5] C. M. Jonker and J. Treur, "Formal analysis of models for the dynamics of trust based on experiences," in *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Springer Berlin Heidelberg, 1999, pp. 221–231.

[6] R. Croson and N. Buchan, "Gender and culture: International experimental evidence from trust games," *American Economic Review*, vol. 89, no. 2, pp. 386–391, 1999.

[7] T. Nomura and S. Takagi, "Exploring effects of educational backgrounds and gender in human-robot interaction," in *2011 International Conference on User Science and Engineering*, 2011, pp. 24–29.

[8] M. Hoogendoorn, S. W. Jaffry, P.-P. Van Maanen, and J. Treur, "Modeling and validation of biased human trust," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 02*. IEEE Computer Society, 2011, pp. 256–263.

[9] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[10] 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.

[11] T. Sanders, K. E. Oleson, D. R. Billings, J. Y. C. Chen, and P. A. Hancock, "A model of human-robot trust: Theoretical model development," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55, 2011, pp. 1432–1436.

[12] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *International Journal of Human-Computer Studies*, vol. 40, no. 1, pp. 153–184, 1994.

[13] G. Hofstede, *Culture's Consequences: International Differences in Work-Related Values*. SAGE Publications, London, 1984.

[14] G. Hofstede, G. J. Hofstede, and M. Minkov, *Cultures and Organizations: Software of the Mind*, 3rd ed. McGraw-Hill Education, 2010.

[15] P. M. Doney, J. P. Cannon, and M. R. Mullen, "Understanding the influence of national culture on the development of trust," *Academy of Management Review*, vol. 23, no. 3, pp. 601–620, 1998.

[16] D. Gefen and T. Heart, "On the need to include national culture as a central issue in e-commerce trust beliefs," *Journal of Global Information Management*, vol. 14, no. 4, pp. 1–30, 2006.

[17] E. Huerta, T. Glandon, and Y. Petrides, "Framing, decision-aid systems, and culture: Exploring influences on fraud investigations," *International Journal of Accounting Information Systems*, vol. 13, no. 4, pp. 316–333, 2012.

[18] S. Rice, K. Kraemer, S. R. Winter, R. Mehta, V. Dunbar, T. G. Rosser, and J. C. Moore, "Passengers from india and the united states have differential opinions about autonomous auto-pilots for commercial flights," *International Journal of Aviation, Aeronautics, and Aerospace*, vol. 1, no. 1, p. 3, 2014.

[19] A. Chaudhuri and L. Gangadharan, "Gender differences in trust and reciprocity," *The University of Auckland, Department of Economics Working Paper Series*, 2003.

[20] N. R. Buchan, R. T. A. Croson, and S. Solnick, "Trust and gender: An examination of behavior and beliefs in the investment game," *Journal of Economic Behavior and Organization*, vol. 68, no. 3-4, pp. 466–476, 2008.

[21] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 442–451, 2008.

[22] T. Koulouri, "Gender differences in navigation dialogues with computer systems," Ph.D. dissertation, Brunel University, School of Information Systems, Computing and Mathematics, 2013.

[23] Amazon, "Amazon mechanical turk," *Amazon Mechanical Turk - Welcome*, 2005, [ONLINE] Available at: https://www.mturk.com/. [Accessed 20 February 2016].

[24] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, 2011.

[25] S. J. Orfanidis, *Introduction to Signal Processing*. Prentice-Hall, Inc., 1995.

[26] C. M. Jonker, J. J. P. Schalken, J. Theeuwes, and J. Treur, "Human experiments in trust dynamics," in *Trust Management: Second International Conference, iTrust 2004*. Springer Berlin Heidelberg, 2004, pp. 206–220.

[27] G. Hofstede, *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*, 2nd ed. SAGE Publications, 2001.

[28] M. P. Haselhuhn, J. A. Kennedy, L. J. Kray, A. B. Van Zant, and M. E. Schweitzer, "Gender differences in trust dynamics: Women trust more than men following a trust violation," *Journal of Experimental Social Psychology*, vol. 56, pp. 104–109, 2015.