Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions

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

Abstract—We developed an experiment to elicit human trust dynamics in human–machine interaction contexts and established a quantitative model of human trust behavior with respect to these contexts. The proposed model describes human trust level as a function of experience, cumulative trust, and expectation bias. We estimated the model parameters using human subject data collected from two experiments. Experiment 1 was designed to excite human trust dynamics using multiple transitions in trust level. Five hundred and eighty-one individuals participated in this experiment. Experiment 2 was an augmentation of Experiment 1 designed to study and incorporate the effects of misses and false alarms in the general model. Three hundred and thirty-three individuals participated in Experiment 2. Beyond considering the dynamics of human trust in automation, this model also characterizes the effects of demographic factors on human trust. In particular, our results show that the effects of national culture and gender on trust are significant. For example, U.S. participants showed a lower trust level and were more sensitive to misses as compared with Indian participants. The resulting trust model is intended for the development of autonomous systems that can respond to changes in human trust level in real time.

Index Terms—Affective computing, autonomous systems, behavioral sciences, cultural differences, data models, human–computer interaction, man–machine systems, stability analysis.

I. INTRODUCTION

THE widespread use of autonomy has improved the quality and efficiency of both safety-critical systems (e.g., nuclear power plants, aircrafts) and devices in daily life (e.g., cars, home appliances). In particular, human trust in autonomous systems is critical to improving the collaboration between humans and such systems [1]. Although various efforts have been made to optimize automated processes, the benefits of automation are lost when humans override these systems due to a fundamental lack of trust [2], [3]. Moreover, accidents may occur due to mistrust [4]. Therefore, trust should be appropriately calibrated to avoid disuse or misuse of automation [5]. One way to potentially overcome these negative effects is to design autonomous systems that can adapt to the human in real time based on the human’s trust level. However, doing this requires predictive, reliable, and quantitative models of human trust behavior.

Researchers have studied trust behavior in human–machine interactions (HMI) and human–computer interactions (HCI) using experimental methods and modeling techniques from social psychology [2], [6], [7]. Some studies focused on analyzing the statistical significance of demographic factors (e.g., age, gender) on trust behaviors [8], [9]. However, while identifying factors that induce changes in trust is a critical step towards characterizing trust behavior; it is alone insufficient for characterizing a quantitative model of this behavior. Moreover, studies have shown that the trust level of humans varies with time due to changing experiences [6], [10] and, as such, any quantitative trust model should be dynamic.

In order to derive a quantitative dynamic model of human trust behavior suitable for HMI contexts, an appropriate experimental design, modeling approach, and model verification are necessary. There is no experimentally verified model for describing the dynamics of human trust level in HMI contexts that: 1) incorporates demographic factors and time-varying experiences; and 2) is built on experiments that elicit multiple transitions in trust level. Existing quantitative models either assume that human trust behavior is fully based on rationale [6] or are nonlinear [10], [11]. While the influence of accumulated effects of past interactions on the future trust level has been modeled in multiagent system contexts, they have not been modeled for independent HMI [11], [12]. Furthermore, existing models of human trust in autonomous systems have not taken into account neither human bias nor attitudes toward the system response bias; here, the system response bias is defined in terms of signal detection theory, i.e., liberal (false-alarm-prone) and conservative (miss-prone) system bias.

Finally, human behavior is highly influenced by one’s surroundings and past experiences [13], [14], which, in turn, are strongly influenced by demographic factors. With the spread of automation across the globe, it is necessary to model human behavior for different demographics. Several types of autonomous systems such as cars, smart thermostats, and tour-guide robots are designed to interact with unspecified users. In these contexts, a model that describes the trust dynamics of a population (general or grouped by demographics) instead of an individual would facilitate the design of such systems. Unfortunately, a generalized model that is suitable for capturing these variations in human trust behavior does not exist in the current literature.

In this paper, we describe the modeling and experimental methods used to capture dynamic changes in human trust,
specifically in an HMI context. This paper significantly expands on our previously published conference paper [15] by including a second experiment that investigates the effects of system error types on trust level. We introduce an improved model structure that explicitly incorporates human bias as an input disturbance to the model. Furthermore, we augment the definition of experience, and the input to the model, to enable the integration of other factors that affect trust, such as system error type. We devise two sets of human subject experiments that elicit multiple dynamic transitions in human trust behavior, and the data collected from these experiments are then used to estimate and validate the parameters of the proposed model. We establish a linear model motivated by literature on computational models for the dynamic variation of human trust [6], [16]; the linearity allows for easier control analysis and synthesis aimed at designing adaptive human–machine interfaces, thus, enabling autonomous systems to respond to human trust variations in real time. Furthermore, we systematically analyze the effects of demographic factors, consisting of national culture [17] and gender, as well as system error type.

This paper is organized as follows. First, we provide background on trust modeling and significant factors that affect human trust. In Section III we describe Experiment 1 along with the development of a generalized model of human trust dynamics; we further examine the effects of national culture and gender on these dynamics. In Section IV we describe Experiment 2 in which we incorporate the effects of misses and false alarms into the model. Afterward, we discuss the implications of the estimated parameter values of the trust model in the context of HMI, followed by concluding statements.

II. BACKGROUND

Comprehensive reviews of the influence of trust in HMI and HCI contexts are provided by Lee and See [5], and Hoff and Bashir [18]. Hoff and Bashir [18] classified trust into three categories: 1) dispositional; 2) situational; and 3) learned [18]. Dispositional trust is based on characteristics of the human. Factors that influence dispositional trust do not vary with time, but they still impact human decision-making during interactions with the autonomous system. Studies have shown differences in trust behavior between people of different cultures, age groups, and personality types [19]–[21]. Situational trust consists of factors that are external to the operator (e.g., task difficulty, potential risks) and those that are internal to the operator (e.g., self-confidence, domain knowledge) [18]. Finally, learned trust is based upon an operator’s overall experience with an autonomous system and influences their initial mindset. During a new interaction with an autonomous system, humans’ experience affects their established trust level. In this paper, we present a dynamic model that can capture variations in human trust level with respect to automation reliability for various demographic groups, thus, capturing both learned and dispositional trust characteristics.

In the remainder of this section, we first introduce studies that have modeled human trust in various contexts. Second, we review literature on determining the effect of automation reliability and system error types (i.e., misses or false alarms) on human trust. These factors influence learned trust. Finally, we review studies on examining dispositional trust factors, specifically gender and national culture.

A. Studies on Human Trust Modeling

Researchers have developed models for predicting human trust based on past experiences, which strongly influence learned trust. Jonker and Treur [6] suggested two types of functions to model trust dynamics: 1) trust update functions; and 2) trust evolution functions. Trust update functions use the current trust level and current experience to update the future trust level, while trust evolution functions map a sequence of trust related events (experiences) to a current trust level. In order to verify the proposed trust dynamics, Jonker et al. [22] conducted follow-up human subject experiments, which presented participants with a sequence of short stories for two scenarios: 1) a photocopying machine; and 2) a travel agency. Each scenario consisted of five positive and five negative stories, and participants reported their trust level after reading each story. The results suggested that trust dynamics are dependent on positive and negative experiences. However, limited by the number of trials (10 trials in each scenario), these studies only induced a single transition in trust level; therefore, the model did not capture the variations in trust dynamics involving multiple transitions.

Some studies have modeled human trust in the context of HMI. Lee and Moray [7] studied changes in human trust level using a simulated semiautomatic juice plant environment. It was observed that the human trust level was affected by the performance of the system, past trust levels, and faults. The authors used an auto-regressive-moving-average-vector analysis to model the input–output relationship of the trust behavior. They later showed that humans use automation when their trust in the automation exceeds their self-confidence [23]. These early efforts demonstrated the effect of situational and learned trusts on the interactions between humans and autonomous systems. However, due to a small sample size (i.e., four to five participants in each group) and a large standard deviation of the data, the accuracy of their model was limited.

Within the simulation context proposed by Moray and his colleagues, Lewandowsky et al. [24] compared trust in automation with trust in human partners in equivalent situations. Similar to the findings of Lee and Moray [23], Lewandowsky et al. [24] identified that faults in auxiliary control actions have a strong effect on trust and self-confidence of the human operator, and the difference between trust and self-confidence is a strong predictor of the human operator’s reliance on automation as well as his/her reliance on human colleagues.

Factors that are significant in predicting trust level may also be dependent on the application context. Sadfaridpour et al. [25] proposed a time-series model for the dynamics of human–robot trust in assembly lines based on the robot performance, human performance, and fault occurrences. More specifically, the performances were quantified by the robot’s working speed and the human’s state of muscle fatigue and recovery. How well the robot met the human’s pace influenced the workload...
and trust perceived by the human. The researchers’ experimental results also suggested that the current trust is mainly dependent on the previous trust if there is no dramatic change in performance.

More recently, elements that are not based on rationale have been incorporated into a human trust model. Li et al. [26] used the structural equation modeling technique to identify the significance of human attitudes and subjective norm on “trusting intentions”. Hoogendoorn et al. [10] developed models with biased experience and/or trust to account for this human behavior. They validated their models using a geographical area classification task and showed that a model with a bias term is capable of estimating trust more accurately than models without an explicit bias. However, their model was nonlinear in trust and experience, rendering it more difficult for analysis than linear models.

B. Effect of Misses and False Alarms in Automation

Automation reliability significantly influences human trust in autonomous systems and, in turn, influences human use of these systems [7], [27]. According to signal detection theory, automation errors can be classified as misses or false alarms; failing to detect the presence of a signal constitutes a miss, and incorrectly alerting humans to an absent stimulus constitutes a false alarm [28]. Existing literature shows that these two types of errors have different effects on human trust in automation. Specifically, these error types affect reliance and compliance to a different degree. On the one hand, reliance is when humans, in the absence of any signal from the system, continue to trust the system and refrain from a response. On the other hand, compliance is exhibited by a human trusting and responding to a signal when the system presents one [29]. An increase in the miss rate reduces reliance, while an increase in false alarms reduces compliance [3], [30]. This distinction is important as it leads the human to react to a signal. For example, a compliance-oriented system (i.e., gives warning when there is a malfunction) increases awareness in humans especially when warnings are spaced close to other indicators [31].

Humans may choose to ignore warnings if they experience high rates of false alarms, which is known as the “cry-wolf” effect [32]. This behavior represents humans’ mistrust of autonomous systems and induces disuse of these systems [33]. Some studies suggest that false alarms cause greater negative effects on human trust in automation as they divert humans’ attention, causing them to monitor unnecessary information [34]. Pervasive false alarms may make humans respond slower or less frequently to similar alerts in future [35], [36]. However, the high false-alarm rate does not appear to negatively impact trust in the context of en route air traffic controller conflict alerts [37]. Indeed, some studies showed contrary results where false alarm prone systems were more trustworthy than miss-prone systems [38], [39]. In addition, there are studies suggesting that false alarms and misses lead to similar effects on trust [40], [41] or that the effect is dependent on humans’ cognitive capabilities [42].

Existing literature shows evident differences in opinions of the effects of misses and false alarms on human trust in automation. In order to resolve these differences, a model for human trust behavior with respect to false alarms and misses is needed. Moreover, the alarm threshold is determined based on the costs associated with each type of error, which means the optimal rate of misses/false alarms varies between systems. Therefore, a model of trust dynamics that connects human trust to autonomous system reliability can help us better understand how reliability-induced trust changes over time. Furthermore, it would allow us to understand how trust recovers with a hit (i.e., system correctly detects the presence of a signal) and/or a correct rejection (i.e., system not alerting the human to an absence of a signal).

C. Demographic Factors That Influence Trust

Autonomous system reliability and error type are external factors that influence learned trust. Apart from experiences accumulated from past interactions with autonomous systems, human trust behavior is also influenced by demographic factors including culture and gender. This is described as dispositional trust and is independent of a specific system or the context of an interaction [18].

Gender differences in trust behavior have been studied thoroughly in economic contexts [8], [43], [44]. Furthermore, some studies have shown gender differences in human–robot interaction contexts and technology adoption behavior [9], [45], [46]. For example, males were more likely to develop trust and positive attitudes toward female robots, while women showed little preference [47]. The attitudes of children toward humanoid robots are also influenced by gender. Tung [48] showed that girls favored humanlike, female robots more than boys did. In addition, females perceived highly automated driving systems as significantly less trustworthy than males did [49].

Values and social norms shared by members of a nation that guide people’s behaviors and beliefs can be defined as the national culture for each country [50]. These factors also have an influence on the cognitive process of trust formation in humans. Therefore, people from different cultures are likely to use different mechanisms to form trust [51] and show particular trust behavioral intentions [52]. To date, only a few studies have examined the effect of national culture on trust in automation. Rice et al. [53] observed that Americans tended to trust less in automated systems as compared to Indians in the context of “auto-pilots.” In another study, Americans were found to trust autonomous (decision-aid) systems less than Mexicans in a fraud investigation scenario [54]. Trust can also be seen as “the willingness to take risk” [55]. Considering the influence of national culture, uncertainty avoidance index (UAI) defined in Hofstede’s six cultural dimensions [50], [56] is relevant to the construct of trust. Uncertainty avoidance tendency has been found to be significant in influencing trust in web design attributes [57], mobile commerce [58], information technology infrastructure [59], and in the context of simulated unmanned air vehicle control [60]. The higher the UAI number for a country, the less likely their people will tolerate uncertainty or risk.
In summary, published quantitative dynamic trust models do not explicitly consider a number of different factors, including the nature of human bias toward the system’s response criteria (i.e., liberal and conservative), demographics, and false alarms and misses in autonomous systems. Moreover, although literature in the area of multiagent systems has analyzed the effects of past experiences on future trust level, this effect needs to be modeled for independent HMI. Therefore, the influence of these factors on human trust dynamics remains unexplored. To address these key gaps, we present two experiments that test the trust factors mentioned above and aid us in establishing a dynamic model of human trust.

III. EXPERIMENT 1

The first experiment was designed to understand human trust dynamics induced by the autonomous system performance, and to identify the effects of humans’ national culture and gender on trust behaviors. The rates of misses and false alarms were controlled; so, participants encountered approximately equal numbers of these two error types. In addition, the order of these two error types was randomized within faulty trials. Therefore, the trust behavior induced by a specific error type was neutralized in Experiment 1.

A. Method

Stimuli and Procedures. The experiment was conducted online, and each participant accessed the study through a computer interface. Participants were told that the experiment was a simplified simulation of driving a car equipped with an obstacle detection sensor. The sensor was based on an image-recognition algorithm that would detect obstacles on the road in front of the car. During each trial, participants’ task was to decide whether or not to trust the algorithm report, based on their previous experience with the algorithm. The instructions informed participants that the image-recognition algorithm used in the sensor was in beta testing.

An experiment session consisted of four initial practice trials followed by 100 trials comprising a sequence of events including stimulus, response, and feedback (see Fig. 1). There were two stimuli: 1) “obstacle detected”; and 2) “clear road,” each having a 50% probability of occurrence. After receiving the stimulus, participants were asked to determine whether they “trusted” or “distrusted” the report provided by the algorithm. The system then gave feedback to the participants on the correctness of their responses (i.e., “correct” and “incorrect”). In order to examine how system reliability influences human trust level, the system was “reliable” in half of the trials and was “faulty” in the remaining half. Here, reliability is defined as the degree to which the algorithm report can be depended on to be accurate. In reliable trials, the algorithm correctly identified the road condition. This meant that “obstacle detected” was a hit, and “clear road” was a correct rejection. Accordingly, it would be marked as “correct” if the participant trusted the report, and “incorrect” if the participant distrusted the report. In faulty trials, there was a 50% probability that the algorithm incorrectly identified the road condition. A report of “obstacle detected” could be a false alarm, and “clear road” could be a miss (see Fig. 2). For the participant, this meant that responding “trust” to a false alarm or a miss would be marked as “incorrect.” On the one hand, we implemented the 100% accuracy condition for reliable trials because it is the ideal performance a sensor can achieve. On the other hand, 50% accuracy for a binary decision would be a pure random chance. Therefore, it should result in the lowest possible trust level that a human has in the simulated sensor.

All the trials in the study (100 in total) were divided into three phases, called “databases,” as shown in Fig. 3. There was a 30-s break before the start of each database. Databases 1 and 2 were used to induce responses to constant system reliability—either reliable or faulty. Database 3 was used to excite all possible dynamics of the participants’ trust responses by switching the accuracy of the algorithm between reliable and faulty according to a pseudorandom binary sequence. Stimuli in each trial were individually randomized for each participant and database. Participants were randomly assigned to one of two groups, which differed in the order of reliable and faulty trials to counterbalance possible ordering effects.

Participants: A total of 581 individuals (ages 20–73) were recruited using Amazon Mechanical Turk [61] to participate in
the study. Among the participants, 340 were males, 235 were females, and 6 did not provide gender information. These participants were randomly assigned to one of the two experimental groups. Participants in Groups 1 and 2 were initially faced with reliable trials and faulty trials, respectively. The participants were paid $0.50 each for their participation in the study. Before starting the study, participants electronically provided their consent. The Institutional Review Board at Purdue University approved the study. We collected participants’ demographic information via a poststudy survey, which included questions about their gender along with the country in which they grew up. The latter is defined as national culture in this study.

Data Processing: To preprocess the collected data, we identified and removed outliers from the dataset. Each participant completed all 100 trials, but they were allowed to skip a trial if they could not make a decision within the given time frame (4 s). We considered excessive “no responses” (i.e., when participants skipped a trial) as well as excessive trust or distrust responses as outliers, determined by the interquartile range (IQR) rule (the 1.5 × IQR rule) [62]. As a result, we removed 63 outliers from the dataset (out of 581 participants), which resulted in 518 valid participants.

To investigate the effects of national culture and gender on human trust, we categorized the collected data into four demographic bins; two were based on nationality: United States (U.S.) and India, and two were based on gender: male and female. Ideally, the selected sample would be representative of the population it came from. However, practically it was not possible to have an equal representation of each demographic group in the collected sample. In order to correct this anomaly in the selection probability of each demographic group in the population, the variables of each bin were adjusted using sampling weights such that each group had an equal representation in the sample population [63]. We calculated sampling weights for each demographic group in all of the bins as follows:

\[
\text{Sampling weight} = \frac{\text{Population percentage}}{\text{Sample percentage}}. \tag{1}
\]

Trust Model Description: For Groups 1 and 2, we computed the probability of trust response for each trial and across all subjects in each of the groups. This probability is defined as the percentage of people in the group who trust the algorithm report. At each trial, for calculating this probability, we assume that the response of each participant is like a Bernoulli trial with “trust” response as success and “distrust” response as failure. Given that for each trial, the responses of all participants are independent from one another, the random variable \(X\), defined as the number of participants responding “trust” on a given trial, has a binomial distribution \(B(k, p)\). The parameter \(k\) is the total number of participants in the bin and the parameter \(p\), binomial proportion, can be estimated using a normal approximation as \(\hat{p} = \frac{x}{n}\). Here \(x\) is the number of successes, i.e., number of trust responses in the given trial across participants. Therefore, at each trial, the probability of trust response can be estimated as \(\hat{p}\) for that trial. The range of estimated probabilities was 0.5 to 1, where 0.5 represented low trust (i.e., the report was perceived as random by the participants; therefore, they responded randomly) and 1 represented high trust. These trust probabilities varied as the decision scenario changed with time and represent the trust level for the sample population. Henceforth, the trust probability will be labeled as trust level \(T(n)\), where \(n \in [1, 100]\) is the trial number. Similarly, we calculated the probability of misses \(M(n)\) and the probability of false alarms \(F(n)\) for each trial, across all subjects, in Groups 1 and 2. For Experiment 1, \(M(n) = F(n)\) and varied from approximately 0 to 0.25, with 0.25 representing faulty trials leading to negative experience and 0 representing reliable trials leading to positive experience. Therefore, we define experience \(E(n)\) as a function of \(M(n)\) and \(F(n)\) given by

\[
E(n) = 1 - [(1 - \beta)M(n) + \beta F(n)]. \tag{2}
\]

Here, \(\beta \in (0, 1)\) is the weighting factor for evaluating the relative effect of misses and false alarms on experience. Beta is the coefficient of the probability of false alarms in the model and, thus, can be called the cry-wolf factor. The higher the value of the cry-wolf factor \(\beta\), the greater the effect of false alarms on the experience, and the lesser the effect of misses on the experience. Since the probability of misses and false alarms was equal for Experiment 1 \([M(n) = F(n) = K(n)]\), (2) reduces to

\[
E(n) = 1 - K(n) \tag{3}
\]

where \(K(n)\) is the probability of a miss or a false alarm. Thus, we obtain the dynamic variation of trust level \(T(n)\) with experience \(E(n)\) for all participants as shown in Fig. 4. In order to reduce noise from the dynamically varying signal \(T(n)\), we used the Savitzky–Golay filter with order 3 and a window of size 5 [64].

Most of the existing human trust models showed trust to be directly related to experience. Jonker and Treur [6] presented change in trust to be directly proportional to the difference of experience and past trust. However, along with experience, we identified the significance of cumulative perception of trust and

Fig. 4. Trust level (probability of trust response) 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. (a) Variation of trust level as a function of trial number. (b) Variation of experience as a function of trial number.
and must lie inside the unit circle, i.e., $|\lambda_{1,2}| < 1$ to guarantee asymptotic stability. Therefore, it is sufficient to prove that

\[
\lambda_1 = 1 - \frac{\alpha + \gamma}{2} - \sqrt{\left(\frac{\alpha - \gamma}{2}\right)^2 + \alpha_\gamma} > -1 \quad (7a)
\]

\[
\lambda_2 = 1 - \frac{\alpha + \gamma}{2} + \sqrt{\left(\frac{\alpha - \gamma}{2}\right)^2 + \alpha_\gamma} < 1 . \quad (7b)
\]

By rearranging and squaring both sides, (7a) and (7b) can be reduced to show that $\forall \gamma \in (0, 1)$, the following hold true:

\[
2 > \alpha + \alpha_e \quad (8a)
\]

\[
0 < \alpha_e + \alpha_b. \quad (8b)
\]

Equations (8a) and (8b) are satisfied if $\alpha_e, \alpha_c, \alpha_b \in (0, 1)$ and $\alpha \in (0, 1)$. Therefore, the trust model (5) is asymptotically stable.

**Remark 1**: The physical interpretation of these bounds on the parameters can be obtained by a closer examination of (4). The parameters $\alpha_e, \alpha_c$ and $\alpha_b$ are weighting factors for each of the terms and should be less than 1 so that the trust level remains stable. The variable $\gamma$ is an exponential weighting factor that belongs to $(0, 1)$. Additionally, $\alpha = \alpha_e + \alpha_c + \alpha_b$ belongs to $(0, 1)$. This ensures that the net coefficient of the term $T(n)$ for calculating $T(n+1)$, i.e., $1 - \alpha$, belongs to $(0, 1)$ and is not negative. Consequently, a higher previous trust level will have a positive influence on the current trust level.

**Proposition 2**: The steady-state values of trust $T_{ss}$ and cumulative trust $C_{T_{ss}}$ for a stable system given by (5) are a weighted average of steady-state experience $E_{ss}$ and expectation bias $B_X$. The weights are proportional to $\alpha_e$ and $\alpha_b$.

**Proof**: By substituting $x(n+1)$ and $x(n)$ with $x_{ss} = [T_{ss} \quad C_{T_{ss}}]^\top$ and $u(n)$ with $u_{ss} = E_{ss}$ in (5), we can solve for the steady-state values $T_{ss}$ and $C_{T_{ss}}$ as follows:

\[
T_{ss} = C_{T_{ss}} = \frac{\alpha_e}{\alpha_e + \alpha_b} E_{ss} + \frac{\alpha_b}{\alpha_e + \alpha_b} B_X . \quad (9)
\]

Here, the subscript $\bullet_{ss}$ represents the steady-state value of the variable.

**Remark 2**: Consider the case when $E_{ss} = 1$, which indicates that the system interacting with the human is consistently accurate. If the expectation bias is less than 1 ($B_X < 1$), the steady-state trust level $T_{ss}$ of the human will be less than 1. Alternatively, consider the case when $E_{ss} = 0$, which indicates that the system interacting with the human is consistently faulty. If $B_X > 0$, the steady-state trust level $T_{ss}$ will also be greater than 0. Therefore, the inclusion of human bias in the proposed model enables us to characterize this important effect on human trust level.

**Parameter Estimation.** For estimating the optimal set of model parameters, we used a nonlinear least squares estimation function *nlgrvest* from MATLAB 2016a. We identified the 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). It is well known that the quality of any empirical
TABLE I
ESTIMATED MEAN PARAMETER VALUES WITH 95% CI FOR ALL PARTICIPANTS AND EACH DEMOGRAPHIC BIN

<table>
<thead>
<tr>
<th>Bin</th>
<th>Experience rate factor $\alpha_x$</th>
<th>Cumulative rate factor $\alpha_c$</th>
<th>Bias rate factor $\alpha_b$</th>
<th>Trust discounting factor $\gamma$</th>
<th>Fit% Grp 1</th>
<th>Fit% Grp 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>$0.2169 \pm 0.0007$</td>
<td>$0.0755 \pm 0.0005$</td>
<td>$0.0428 \pm 0.0004$</td>
<td>$0.1148 \pm 0.0012$</td>
<td>95.7138</td>
<td>92.5567</td>
</tr>
<tr>
<td>US</td>
<td>$0.2157 \pm 0.0007$</td>
<td>$0.0635 \pm 0.0007$</td>
<td>$0.0394 \pm 0.0004$</td>
<td>$0.1270 \pm 0.0029$</td>
<td>94.5923</td>
<td>87.5857</td>
</tr>
<tr>
<td>India</td>
<td>$0.2177 \pm 0.0010$</td>
<td>$0.0996 \pm 0.0011$</td>
<td>$0.0515 \pm 0.0007$</td>
<td>$0.0942 \pm 0.0008$</td>
<td>91.9665</td>
<td>90.1372</td>
</tr>
<tr>
<td>Female</td>
<td>$0.2277 \pm 0.0009$</td>
<td>$0.0783 \pm 0.0007$</td>
<td>$0.0373 \pm 0.0005$</td>
<td>$0.1042 \pm 0.0017$</td>
<td>91.4777</td>
<td>89.1182</td>
</tr>
<tr>
<td>Male</td>
<td>$0.2085 \pm 0.0009$</td>
<td>$0.0817 \pm 0.0007$</td>
<td>$0.0491 \pm 0.0005$</td>
<td>$0.1375 \pm 0.0018$</td>
<td>93.9327</td>
<td>89.1748</td>
</tr>
</tbody>
</table>

Fig. 5. 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. (a) Group 1: $R^2 = 95.74\%$. (b) Group 2: $R^2 = 92.53\%$.

parameter estimation is dependent on the data itself. A sample of human subject data cannot completely represent the human population, and the derived inferences may vary based on the selected sample. Therefore, in order to verify the robustness of the parameter estimation relative to the sample selection, we iterated the estimation 1000 times, with each iteration using a new, randomly selected subset of data representing 90% of the total datasets for all participants and each demographic bin. There was less than 2.5% error in the estimated parameter values caused by the variation in sample selection for a 95% confidence interval (CI) (see Table I), signifying a robust estimation.

B. Results

In order to verify whether our proposed model of trust level is valid, we estimated the model parameters for a general population, which included all 518 valid participants in our experiment. The fit between the trust model and the experimental data is shown in Fig. 5. Table I shows the optimal parameter values and the goodness of fit between the data and the model calculated using R-squared. The goodness of fit was 95.71% and 92.56% for all participants in Groups 1 and 2, respectively. Note that all the estimated parameter values satisfy the stability criteria defined in Proposition 1.

We observed that participants from different demographic groups required different amounts of time to adapt to changes in the system performance and attained different steady-state trust levels. In order to analyze these differences, we simulated the step response of each parameterized model. A sample step response for all participants with expectation bias $B_X = 0$ is shown in Fig. 6. The calculated rise time for the step response (see Table II) is an indicator of the rate of change of the trust dynamics. Rise time is defined as the time required for the response to increase from 10% to 90% of its final value. Therefore, a longer rise time implies slower trust dynamics.

Fig. 7 shows the experimentally obtained trust level and the predicted trust level of participants grouped by their national culture. Upon visual inspection, the U.S. participants trusted the system report less during the trials in Database 3, than during trials in Databases 1 and 2, in which the accuracy of the algorithm was switched between reliable and faulty; see Fig. 7(a) and (c). Moreover, in response to changes in system reliability, the trust level of U.S. participants changed at a faster rate, and approached an overall lower level than that of Indian participants. These observations are supported by the calculated rise time of the models (see Table II). The rise time of the state $T$ for Indian participants is 53.8% higher than that of U.S. participants. This implies that Indian participants’ trust level increased or decreased more slowly than that of U.S. participants after the system performance changes. Additionally, the rise time of the state $C_T$ for U.S. participants is 29.4% shorter than that of Indian participants, which implies that their cumulative trust changed...
Fig. 7. Participants grouped by national culture. 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) U.S. Group 1: $R^2 = 94.51\%$. (b) India Group 1: $R^2 = 92.00\%$. (c) U.S. Group 2: $R^2 = 87.56\%$. (d) India Group 2: $R^2 = 90.08\%$.

Relatively faster. This observation can also be attributed to the trust discounting factor $\gamma$, which is 34.8% larger for U.S. participants, indicating that U.S. participants relied on their recent trust level and experience more as compared to Indian participants.

Fig. 8 shows the experimentally obtained trust level and the prediction of participants grouped by their gender. The plots show that male participants exhibited greater trust in the system than female participants, especially when the system did not perform well [see Fig. 8(b) and (d)]. However, the trust level of female participants changed more rapidly than that of male participants. Similarly, when comparing the step responses, the rise time of state $T$ for male participants is 15.4% longer than that of female participants, implying that the trust level of male participants changed more slowly than that of female participants. Furthermore, the rise time of the state $C_T$ for male participants is 14.8% shorter than that of female participants, which implies that their cumulative trust changed relatively faster. This observation can also be attributed to the trust discounting factor $\gamma$, which is 32.0% larger for male participants, indicating that they relied on their recent trust level more as compared to female participants.

Based upon the high fit percentages achieved between the model and experimental data after parameter estimation, these results suggest that human trust in autonomous systems can be modeled as a function of their experience (which varies with system performance), cumulative trust, and expectation bias. Moreover, the estimated model parameters capture the effects of national culture and gender on trust behaviors.

IV. EXPERIMENT 2

As an extension of Experiment 1, we designed Experiment 2 to conduct an in-depth study on the effects of misses and false alarms on participants’ trust levels. In this experiment,
Fig. 9. Trust level (probability of trust response) for all participants and the probability of misses/false alarms that affect the experience. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks. Participants showed trust in reliable trials and distrust in faulty trials. (a) Variation of trust level as a function of trial number. (b) Variation of misses/false alarms as a function of trial number.

we present participants with trials containing 100% of misses and 100% of false alarms, unlike the 50–50 split used in Experiment 1 (see Fig. 9).

A. Method

We followed the same methodologies from Experiment 1 in terms of data collection, data processing, and modeling. We revised the stimuli to elicit trust reactions in response to misses and false alarms and analyzed the resulting data that were collected. We then expanded the general trust model to incorporate the effects of misses and false alarms.

Stimuli and Procedures. In comparison to Experiment 1, the only additional factor incorporated into Experiment 2 was the error type during faulty trials. More specifically, we manipulated the probability of misses and false alarms in faulty trials. In Experiment 1, a system error was equally probable to be a miss or a false alarm in each faulty trial. In Experiment 2, we examined the following three conditions:

1) an error was always a miss in a session of faulty trials;
2) an error was always a false alarm in a session of faulty trials;
3) an error was equally probable to be a miss or a false alarm in a session of faulty trials.

Fig. 10 shows the condition and trial orders in each database. Participants were randomly assigned to one of two groups in the interest of testing whether the experience of misses or false alarms affects the other.

Participants: A total of 333 individuals (ages 19–74) participated in Experiment 2. Among the participants, 171 were males, 158 were females, and 4 did not provide gender information. These participants were randomly assigned to one of the two experimental groups. The recruitment procedure and the survey used to collect demographic information were the same as those in Experiment 1.

Data processing: We used the IQR rule as introduced in Experiment 1 to identify and remove outliers. The procedure resulted in 293 valid datasets (out of a total of 333 participants) to be analyzed.

Parameter Estimation: Using the data collected in Experiment 2, we estimated the cry-wolf factor $\beta$ by setting all other factors (experience rate factor $\alpha_e$, cumulative rate factor $\alpha_c$, bias rate factor $\alpha_b$, and trust discounting factor $\gamma$) to the values estimated in Experiment 1. The robustness of the estimated value of $\beta$ was verified by 1000 iterative estimations, 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% CI were less than 2.5%. Table III shows the parameter values and the goodness of fit.

B. Results

We first investigated whether the system error type (i.e., miss and false alarm) affects the trust dynamics of the general population, which included all 293 valid participants in the experiment. Fig. 11 shows the experimental trust level compared against the model. Participants responded differently to misses and false alarms, and in some cases, the experience of one error type affected later responses to the other error type. The proposed trust model was able to predict the trust dynamics while taking into account the rate of misses and false alarms. The goodness of fit was measured using the R-squared value of the data; the results
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

were 91.76% and 91.20% for all participants in Groups 1 and 2, respectively.

The results suggest an interaction effect between risk-taking behavior and demographic factors on trust. Fig. 12 shows the experimentally obtained trust level and model predictions for participants grouped by their national culture. U.S. participants trusted less than Indian participants when encountering system misses [see Fig. 12(a) and (b)]. Moreover, U.S. participants trusted less in miss-prone trials than false alarm prone trials regardless of whether they encountered false alarms first or not [see Fig. 12(a) and (c)]. By contrast, Indian participants trusted less in miss-prone trials than false alarm prone trials regardless of whether they encountered false alarms first or not [see Fig. 12(a) and (c)]. By contrast, Indian participants trusted less in miss-prone trials than false alarm prone trials when they encountered misses first [see Fig. 12(b)]; their trust level in miss-prone trials decreased less if they encountered system false alarms prior to misses [see Fig. 12(d)]. The cry-wolf factor $\beta$ of the model is 48.3% larger for Indian participants than for U.S. participants. The larger the value of $\beta$, the weaker the negative effect of misses on trust, indicating that misses have a stronger negative effect on trust for U.S. participants as compared with Indian participants.

We also observed gender differences in response to system misses and false alarms. Fig. 13 shows the experimental trust level and the prediction of participants grouped by their gender. Male participants trusted less in miss-prone trials than female participants if they had not encountered system false alarms first [compare Fig. 13(a) and (b)]. However, if participants encountered false alarms first, females reached a lower trust level than males [compare Fig. 13(c) and (d)]. In general, male participants were more sensitive to system misses. The cry-wolf factor $\beta$ of the trust model supports this observation; $\beta$ is 18.0% larger for female participants than for male participants, which implies that misses have a stronger negative effect on trust for male participants as compared with female participants.

V. DISCUSSION

In this section, we provide a more in-depth discussion of the main results of the two experiments. The two experiments presented in this paper elicited the variation of a human’s trust response to system reliability. Participants attained a high trust level in reliable trials and a 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 trust dynamics were modeled based on past behavioral
Moreover, in some cases [e.g., Fig. 13(a)] the trust response still increased or decreased near this newly attained level in both reliable and faulty trials. This finding was contrary to Jonker et al. [22] 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. However, we found this to be less than the required number of trials to reach a new trust level.

The aggregated trust response and the trust model enhanced our understanding of dispositional trust and learned trust in autonomous systems. Participants from the U.S. exhibited a lower trust level than Indian participants. This is consistent with the findings from Rice et al. [53] and Huerta et al. [54] that Americans trust autonomous systems less than Indians and Mexicans, respectively. Moreover, system misses induced stronger distrust in U.S. participants than in Indian participants, suggesting that U.S. participants are less willing to take risks. This agrees with the smaller UAI of Indian culture as compared to that of U.S. culture (40 versus 46) [17], where the literature demonstrated that humans from higher uncertainty avoidance cultures are less likely to trust or implement new technology [58], [59].

Regarding gender, male participants appeared to trust more than female participants, especially when the system was not reliable. This is supported by Feldhutter et al. [49], but is contrary to the findings of Haselhuhn et al. [65], which showed that women’s trust decreases less than men after transgressions as they prefer to maintain interpersonal relationships. These results highlight that the dynamics of human trust behavior in HMI contexts is different from interpersonal trust behavior between humans, thus, creating a need for human trust models in HMI contexts. Additionally, the variation in trust responses of female participants was noticeably higher than that in male participants. This variation indicates that the female participants have diverse perceptions of autonomous systems and, therefore, other factors such as personality and expertise should be investigated in future studies. Finally, there were gender differences in the responses to misses and false alarms as discussed in Section IV-B. Along with the observations of U.S. and Indian participants, demographic effects can partially explain the inconsistency between previously published results on the effects of system error type on human trust.

VI. CONCLUSIONS

We developed a study composed of two experiments to elicit a dynamic change in human trust with respect to HMI contexts. Furthermore, we established a quantitative trust model, motivated by literature on computational models and parameterized using human subject data. This model was verified using data collected from more than 800 participants and has a prediction accuracy higher than 92% for the general population. We introduced the effect of cumulative trust, expectation bias, and misses/false alarms, to accurately capture human trust dynamics during HMI.

It has been established that trust plays an important role in human–system interactions. Therefore, to establish collaborative interactions between humans and autonomous systems, it is essential to adapt the human–system interaction based on the human’s trust level. This, in turn, requires autonomous
systems to utilize quantitative models of dynamic human trust behavior. Existing human trust models are typically nonlinear or predict trust solely based on experiences. Moreover, others were developed using experimental data in which the stimulus—a transition in system performance—occurred only once in each experiment. Therefore, their ability to predict trust variations is limited. We addressed this gap by identifying the significance of cumulative trust and expectation bias through experiments that elicited multiple dynamic transitions in human trust, and then incorporated these two variables in the proposed linear model.

In addition to proposing a general trust model structure, we characterized the effects of both dispositional and learned trust factors, specifically national culture, gender and system error type, using estimated model parameters. We also characterized the effects of misses and false alarms on the dynamics of human trust behavior and compared differences between demographics. We found that system misses induce a stronger distrust in U.S. participants than that in Indian participants and have a stronger negative effect on trust for male participants than that for female participants. While the proposed model is representative of a population of individuals rather than trained to a specific human, such a model could be used to design machines that are required to interact with unspecified users grouped by demographics.

One limitation of this study is that a computer-based interface system was used in the experiment, and therefore, the ecological validity could be improved. Future work will involve conducting experiments in real-life settings. The model could also be generalized for use in a wider range of domains by expanding the definition of experience to incorporate other significant factors beyond the probability of misses and false alarms, such as system transparency and the level of automation. Other extensions of the work will include consideration of additional demographicson and validation of the model in other HMI contexts.

REFERENCES


HU et al.: COMPUTATIONAL MODELING OF THE DYNAMICS OF HUMAN TRUST DURING HUMAN–MACHINE INTERACTIONS


Wan-Lin Hu received the B.S. degree in bio-industrial mechatronics engineering and the M.S. degree in electrical engineering from National Taiwan University, Taipei City, Taiwan, in 2006 and 2008, respectively. She received the Ph.D. degree in mechanical engineering from Purdue University, West Lafayette, IN, USA, in 2017. She is a Postdoctoral Scholar with the SLAC National Accelerator Laboratory, Stanford University, Stanford, CA, USA. Her research interests include human–machine interaction, psychophysiology, decision-making, and design methodology.

Kumar Akash received the B. Tech. degree in mechanical engineering (ME) from the Indian Institute of Technology Delhi, New Delhi, India, in 2015, and the M.S. degree in ME from Purdue University, West Lafayette, IN, USA, in 2018. He is working toward the Ph.D. degree at the School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA. His research interests focus on dynamic modeling and control of human behavior in human–machine interactions, brain–computer interfaces, and machine learning.

Tahira Reid received the B.S. and M.S. degrees in mechanical engineering (ME) from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2000 and 2004, respectively. She received the Ph.D. degree in design science from the University of Michigan, Ann Arbor, MI, USA, where ME and psychology were her focus areas.

She is an Associate Professor with the School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA. Her research interests focus on the quantification and integration of human-centered considerations in engineering, the design process, and human–machine systems.

Neera Jain received the S.B. degree in mechanical engineering (ME) from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2006, and the M.S. and Ph.D. degrees, also in ME, from the University of Illinois at Urbana-Champaign in 2009 and 2013, respectively.

She is an Assistant Professor with the School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA. Her research interests include dynamic modeling and control theory with applications to human–machine interactions, thermodynamic systems, and advanced manufacturing.