Modeling Trust Dynamics in Human-robot Teaming: A Bayesian Inference Approach

Yaohui Guo
University of Michigan, Ann Arbor
Ann Arbor, MI 48105, USA
yaohuig@umich.edu

Chongjie Zhang
Tsinghua University
Beijing, 100084, P. R. China
chongjie@tsinghua.edu.cn

X. Jessie Yang
University of Michigan, Ann Arbor
Ann Arbor, MI 48105, USA
xijyang@umich.edu

CHI'20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3383007

Abstract
In this work, we proposed a personalized trust predictor for modeling trust dynamics in human-robot teaming. The proposed method models trust by a Beta distribution to capture the three properties of trust dynamics, which takes the performance-induced positive attitude and negative attitude as parameters. The model learns the prior distribution of the parameters from a training dataset, and estimates the posterior distribution based on a short training session and occasionally reported trust feedback. The experiments showed that the proposed method accurately predicted people's trust dynamics, achieving a root mean square (RMS) of 0.0724 out of 1.

Author Keywords
Trust in automation; Trust in autonomy; Human-automation interaction; Human-robot interaction; Human-robot teaming

CCS Concepts
•Human-centered computing → HCI theory, concepts and models;

Introduction
Advances in robotics enable robots to assist humans in a variety of fields, including transportation, healthcare, and manufacturing. The human-robot team's success relies on the ability of both the human and the robotic agents to col-

Problem statement

Goal:
Predict the human agent's moment-to-moment trust based on his or her interaction history with the robot and occasionally reported trust feedback.

Input:
Robot's performance history, human agent's reported trust feedback during the training session and occasionally reported trust feedback.

Output:
Moment-to-moment trust.

laborate with each other. Just like human-human teaming, to ensure effective human-robot teaming, appropriate trust has to be established between the human and the robot [1, 3, 4, 7].

Despite the research efforts on trust in automation/autonomy over the past three decades, one major research gap remains: The majority of prior literature adopted a "snapshot" view of trust and typically measured trust once through questionnaires at the end of an experiment. More than two dozen factors have been identified to influence one's (snapshot) trust in autonomy, including individual factors such as culture and age [9, 12, 2], system factors such as reliability and level of automation [10, 11, 14, 15], and environmental factors such as multi-tasking requirement [18]. This "snapshot" view, however, does not acknowledge that trust is a time-variant variable that can strengthen and decay over time. With few exceptions (e.g. [6, 13, 17, 16]), we have little understanding of the temporal dynamics of trust formation and evolution, nor of how trust strengthens or decays over time as a result of moment-to-moment interactions in human-agent teams.

In the present study, we proposed a Bayesian personalized trust predictor to model trust dynamics in human-robot teaming. The proposed method models trust as a modified Beta distribution to capture the three properties of trust dynamics, which takes the performance-induced positive attitude and negative attitude as parameters. The model learns the prior distribution of the parameters from a training dataset, and estimates the posterior distribution based on a short training session and occasionally reported trust feedback. Using an existing dataset collected by by Yang et al. in [16], we showed that the proposed method accurately predicted human operators' trust in a robotic agent.

Problem statement

This work is aimed to build a personalized trust predictor for estimating a human agent's moment-to-moment trust during human-robot interaction. The predictor is able to estimate the human agent's moment-to-moment trust only based on some occasionally trust feedback after a short training session.

Let's consider an example where an assistant robot is designed to work with human operators to perform a series of tasks. We denote the robot's performance on the i^{th} task as $p_i \in \{0, 1\}$, where $p_i = 1$ indicates a success while $p_i = 0$ indicates a failure. The reliability of the robot, $r \in [0, 1]$, is defined as the probability that the robot can succeed the task. Here we assume the robot has the same reliability while working with one operator, but its reliability may vary between operators. At time i, after observing the robot's performance p_i, the operator will update his/her trust $t_i \in [0, 1]$ in the robot according to the performance history $\{p_1, p_2, ..., p_i\}$, where $t_i = 1$ means the operator completely trusts the robot and $t_i = 0$ means the operator does not trust the robot at all.

Suppose the robot has previously been trained with k human operators and completed n tasks with each operator. Each operator provided his/her trust feedback t_i at the end of each task i. Therefore, the trust history $T^j = \{t_1^j, ..., t_n^j\}$ and the robot's performance history $P^j = \{p_1^j, ..., p_n^j\}$ are fully available, $j = 1, 2, ..., k$. Now a new operator will work with the robot for the first time: the operator will be trained working with the robot for the first l tasks and during this training s/he will report his/her trust after each task; after this training session, the operator will continue working with the robot, but s/he can choose to or not to provide his/her trust feedback after each task.

The objective is defined as the following: after the new operator finishes the m^{th} task, given the robot's performance history $P_m = \{p_i | i = 1, 2, 3, ..., m\}$, trust history of the training session $T_m^t = \{t_i | i = 1, 2, 3, ..., l\}$, occasionally reported trust $T_m^o = \{t_i | i \in O_m, O_m \subset \{l+1, l+2, ..., m-1\}\}$, and the data from the k operators T^j and P^j, $j = 1, 2, ..., k$, predict the current trust t_m. Here O_m is an indicator set: $O_m = O_{m-1} \cup \{m-1\}$ if the user choose to report his/her trust after the $m-1$th task, otherwise $O_m = O_{m-1}$. We define trust history at time m as $T_m = T_m^o \cup T_m^t$.

Proposed model

We proposed to model trust dynamics using the Beta distribution. We showed that the proposed model can reflect the three properties of trust dynamics.

Personalized trust prediction model

Based on the related studies, a desired trust model should have the following three properties:

1. Trust at the present moment is determined by trust at the previous moment [5];
2. Negative experience had more influence on trust than positive experience and a single automation failure led to immediately decrease of trust [8];
3. Human operators' trust in automation would stabilize over repeated interaction with an automated technology [16].

To reflect these properties, we proposed to use Bayesian inference with the Beta distribution to predict human trust. We propose that after the robot completing the i^{th} task, the operator's temporary trust t_i follows a Beta distribution:

$$t_i \sim Beta(\alpha_i, \beta_i) \tag{1}$$

The predicted trust $\bar{t}_i$ is given by the mean of the distribution

$$\bar{t}_i = E(t_i) = \frac{\alpha_i}{\alpha_i + \beta_i} \tag{2}$$

α_i and β_i are updated by

$$\alpha_i = \begin{cases} \alpha_{i-1} + w^s & \text{, if } p_i = 1 \\ \alpha_{i-1} & \text{, if } p_i = 0 \end{cases}$$
$$\beta_i = \begin{cases} \beta_{i-1} + w^f & \text{, if } p_i = 0 \\ \beta_{i-1} & \text{, if } p_i = 1 \end{cases} \tag{3}$$

again p_i is the performance of the robot on the i^{th} task. Here α_i and β_i correspond to the operator's positive and negative experience with the robot, which can be viewed as the negative and positive attitude the operator gained from the interaction experience respectively. w^s and w^f are the gains of the positive attitude and negative attitude at each task respectively, where the superscript s stands for success and f stands for failure.

Next we show that the model features the three properties of trust dynamics. First, it is clear in Eq. (3) that the present trust is determined by the previous trust, so the first property is satisfied. Second, we calculate the difference between the increase of trust cause by automation success and decrease of trust caused by automation failure at time i:

$$\begin{aligned} &(\bar{t}_i|_{p_i=1} - \bar{t}_{i-1}) - (\bar{t}_{i-1} - \bar{t}_i|_{p_i=0}) \\ =&\frac{1}{D}\left(\frac{w^s \beta_{i-1}}{D + w^s} - \frac{w^f \alpha_{i-1}}{D + w^f}\right) \end{aligned} \tag{4}$$

where $D = \alpha_{i-1} + \beta_{i-1}$. If α_{i-1} and β_{i-1} are close, then Eq. (4) indicates that an automation failure will lead to a larger trust change compared to an automation success when $w^f > w^s$. More precisely, when $\frac{\alpha}{\beta} > \frac{w^s D + w^s w^f}{w^f D + w^s w^f}$, the automation failures will have a larger impact. So the second property will be satisfied in most cases when the value of w^s and w^f are appropriately chosen. Finally, for

Model inference

The posterior is estimated via MAE, and the prior is learned via MLE from the data of the other k operators.

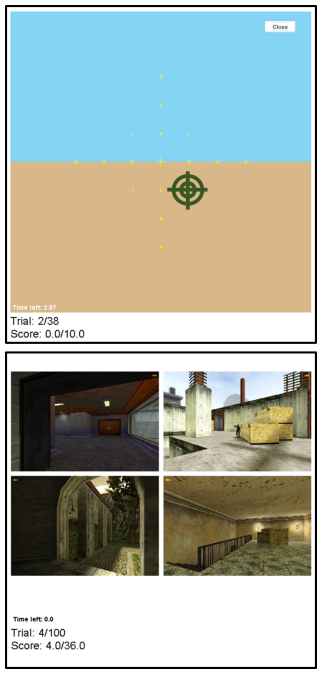


Figure 1: Dual-task environment in the simulation testbed. The two images show displays from the simulation testbed for the tracking (top) and detection (bottom) tasks respectively. Participants could access only one of the two displays at a time, and could switch between them.

the stabilization property, let's suppose the robot has a constant reliability r. After n tasks, the robot accomplishes n^s tasks and fails n^f tasks. Then

$$t_n \sim Beta(\alpha_0 + n^s w^s, \beta_0 + n^f w^f) \tag{5}$$

When $n \to \infty$, t_n will be a point mass distribution centered at

$$\frac{\alpha_0 + n^s w^s}{\alpha_0 + \beta_0 + n^f w^f + n^s w^s} = \frac{r w^s}{r w^s + (1-r) w^f} \tag{6}$$

which means trust stabilizes with repeated tasks. Therefore, all the three properties of trust dynamics are satisfied in the proposed model.

Now we discuss how to infer the model parameters. Given robot performance history $\{p_1, p_2, ..., p_n\}$, trust $\{t_1, t_2, ..., t_n\}$ can be totally determined by the parameter set

$$\theta = \{\alpha_0, \beta_0, w^s, w^f \tag{7}$$

So to personalize the trust model for a certain operator is to find the best θ for him. Here we use maximum a posterior estimation (MAP) to estimate θ, which is to maximize the posterior of θ given the robot performance P_m, trust history T_m and robot reliability r. Because

$$P(\theta \mid P_m, T_m, r) \propto \prod_{t_i \in T_m} Beta(t_i; \alpha_i, \beta_i) \cdot P(\theta) \tag{8}$$

we have

$$\theta = \underset{\theta}{\text{argmax}} \sum_{t_i \in T_m} \log(Beta(t_i; \alpha_i, \beta_i)) + \log P(\theta) \tag{9}$$

The above equation shows that θ will be updated only when the operator produces a new trust feedback. The prior $P(\theta)$ can be learned by maximum likelihood estimation (MLE) from the data of other operators, namely T^j and P^j, $j = 1, 2, ..., k$

$$\begin{aligned} \theta &= \underset{\theta}{\text{argmax}} \prod_{j=1}^{k} P(T^j, P^j \mid \theta) \\ &= \underset{\theta}{\text{argmax}} \prod_{j=1}^{k} \prod_{i=1}^{n} Beta(t_i^j; \alpha_i^j, \beta_i^j) \end{aligned} \tag{10}$$

Experiments

In the experiment, we tested our trust model on a dataset where the participants were asked to report their trust towards an automated threat detector. We analyzed the training and prediction results.

Dataset

In this work, we utilized the dataset collected by Yang et al. [16]. Participants in the experiment performed a simulated surveillance task consisting of a tracking task and a detection task (Fig. 1). For the tracking task, participants controlled a joystick and moved the green circle to the center of the display as close as possible. Meanwhile, participants were asked to detect whether there was a potential threat in four images. Participants were able to access only one task at any time and had to switch between the tracking task and the detection task. There was an imperfect threat detector to assist human operators in detecting the threat. While two kinds of detectors were introduced in [16], we only consider the cases where a binary detector was used. The system reliability of the threat detector was set as 70%, 80%, and 90%. Each participant had 100 trials with each trial lasting 10 seconds. After each trial, participants reported their perceived automation reliability, trust in automation, and confidence. Here we only use the infor-

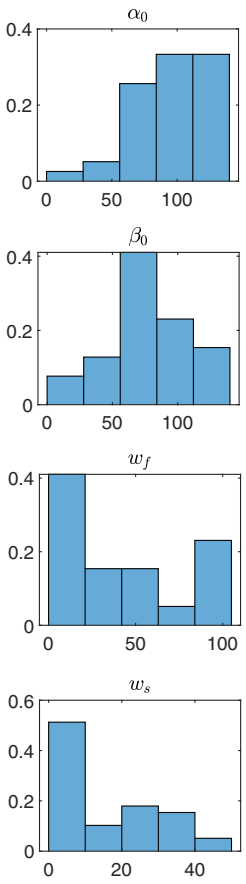


Figure 3: Learned distribution of w^s, w^f, α_0, and β_0.

mation of the operator's trust feedback and the detector's performance according to the problem statement.

Experiments

There were 39 participants in total who worked with a binary detector. Due to this limited number of data points, we used the leave-one-out method to evaluate the proposed model: in each run, one participant's data was picked out as the testing data while the other 38 participants' data were used as the training data. During the training, a human agent's trust history and the detector's performance history were fully available. During testing, after the m^{th} trial, where $m > l$, input to the predictor included the trust history of the training session $T_m^t = \{t_i | i = 1, 2, 3, ..., l\}$ and the human agent's occasionally reported trust feedback $T_m^o = \{t_i | i = l + q, l + 2q, l + 3q, ..., i < m\}$. Here we assume the operator reported his/her trust in every q trials after the l personalized training trials. In this section we set $l = 10$ and $q = 10$. How different l and q affect the prediction will be discussed later.

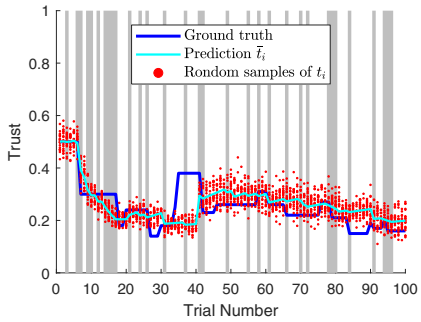


Figure 2: Trust prediction for the first operator. The first 10 trust feedback are given, and further trust feedback is given every other 10 trials. The gray zone indicates a failure, while the white zone means success. The cyan curve is the predicted trust.

Training

The distributions of $\alpha_0, \beta_0, w_s, w_f$ learned in the training are shown in Fig. 3. It is clear that α has a larger mean than β, which indicates that the participants in the experiment generally have a more positive attitude towards the detector. Also, w_f's mean is larger than w_s's mean, so in the experiment a detection failure would change trust more compared to a detection success.

Prediction

We evaluated the root mean square error (RMS) of the proposed method. the RMS of the proposed method is 0.0724. Fig. 2 illustrates the prediction result for the first operator. It shows that the predicted trust successfully captures the operator's trust dynamics.

Discussion

In this section, we discuss the three types of trust dynamics and how trust report frequency affected the prediction results.

Three types of trust dynamics

In the experiments, we found that the operators' trust dynamics can be categorized into three types: the rational agent whose trust dynamics can be modelled accurately by Bayesian inference (Fig. 4a), the oscillator whose trust changes abruptly (Fig. 4b), and the disbeliever whose trust is constantly low no matter how capable an autonomous agent is (Fig. 4c).

Above are the three typical types of trust behavior observed in the dataset, but there is no hard boundary between them. For example, fluctuation can be found in many participants, while their overall trust feedback are reasonable.

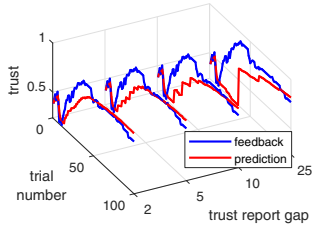
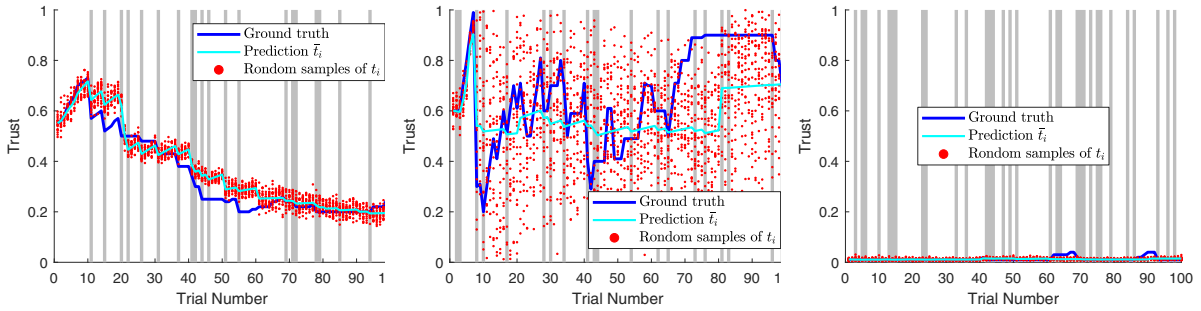


Figure 5: Prediction results under different trust report gaps. Number of personalized steps is 10.



(a) Rational decision maker **(b)** Oscillator **(c)** Disbeliever

Figure 4: Three different types of trust behaviors

Trust report frequency

We examined how the trust report frequency affected the prediction results. One human participant's data is used as an illustration. In Fig. 5, the prediction results were produced by setting $l = 10$ and $q = 2, 5, 10, 25$ respectively. It is clearly shown that the smaller the gap is, the better the prediction result is. Since after the training session the model parameters will only be updated when a new trust feedback is provided, there are "jumps" on the prediction curve whenever the human operator chooses to report his/her trust. For the operator in Fig. 5, a larger trust report gap, such as 25, will make the prediction accuracy unacceptable, while asking the operator for more trust feedback will disturb the operator, so it is important to find a balanced trust report gap to maximize the human-robot team performance.

Conclusion

In this work, we developed a personalized trust predictor based on the Bayesian framework. Motivated by the three properties of trust dynamics from related literature, we proposed to model trust by a Beta distribution. We evaluated the model using an existing dataset and showed the proposed model achieved a RMS of 0.0724. Moreover, as the model only had four parameters, it can be inferred fast and thus used for real-time tasks.

REFERENCES

[1] E. J. de Visser, R. Pak, and T. H. Shaw. From automation to autonomy: the importance of trust repair in human-machine interaction. *Ergonomics*, 133(1):1–19, apr 2018.

[2] N. Ezer, A. D. Fisk, and W. A. Rogers. Age-Related Differences in Reliance Behavior Attributable to Costs Within a Human-Decision Aid System. *Human Factors*, 50(6):853–863, dec 2008.

[3] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, 53(5):517–527, sep 2011.

[4] K. A. Hoff and M. Bashir. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57(3):407–434, Apr. 2015.

[5] J. Lee and N. Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.

[6] J. D. Lee and N. Moray. Trust control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.

[7] J. D. Lee and K. A. See. Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[8] D. Manzey, J. Reichenbach, and L. Onnasch. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6(1):57–87, 2012.

[9] S. E. McBride, W. A. Rogers, and A. D. Fisk. Understanding the effect of workload on automation use for younger and older adults. *Human Factors*, 53(6):672–686, nov 2011.

[10] R. Parasuraman and V. Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2):230–253, 1997.

[11] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000.

[12] P. P. Rau, Y. Li, and D. Li. Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior*, 25(2):587–595, mar 2009.

[13] C. Wang, C. Zhang, and X. J. Yang. Automation reliability and trust: A Bayesian inference approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1):202–206, 2018.

[14] C. D. Wickens and S. R. Dixon. The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, 8(3):201–212, May 2007.

[15] C. D. Wickens, S. Rice, D. Keller, S. Hutchins, J. Hughes, and K. Clayton. False Alerts in Air Traffic Control Conflict Alerting System: Is There a "Cry Wolf" Effect? *Human Factors*, 51(4):446–462, 2009.

[16] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah. Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 408–416. IEEE, 2017.

[17] X. J. Yang, C. D. Wickens, and K. Hölttä-Otto. How users adjust trust in automation: Contrast effect and hindsight bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1):196–200, 2016.

[18] M. Y. Zhang and X. J. Yang. Evaluating effects of workload on trust in automation, attention allocation and dual-task performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1):1799–1803, 2017.