TIP: A *T*rust *I*nference and *P*ropagation Model in Multi-Human Multi-Robot Teams

Yaohui Guo
University of Michigan
Ann Arbor, Michigan, USA
yaohuig@Umich.edu

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

Cong Shi
University of Michigan
Ann Arbor, Michigan, USA
shicong@umich.edu

ABSTRACT

Trust has been identified as a central factor for effective human-robot teaming. Existing literature on trust modeling predominantly focuses on dyadic human-autonomy teams where one human agent interacts with one robot. There is little, if not no, research on trust modeling in teams consisting of multiple human agents and multiple robotic agents. To fill this research gap, we present the trust inference and propagation (TIP) model for trust modeling in multi-human multi-robot teams. We assert that in a multi-human multi-robot team, there exist two types of experiences that any human agent has with any robot: direct and indirect experiences. The TIP model presents a novel mathematical framework that explicitly accounts for both types of experiences. To evaluate the model, we conducted a human-subject experiment with 15 pairs of participants ($N = 30$). Each pair performed a search and detection task with two drones. Results show that our TIP model successfully captured the underlying trust dynamics and significantly outperformed a baseline model. To the best of our knowledge, the TIP model is the first mathematical framework for computational trust modeling in multi-human multi-robot teams.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Computer systems organization → Robotics**.

KEYWORDS

team of teams, multi-operator multi-autonomy (MOMA)

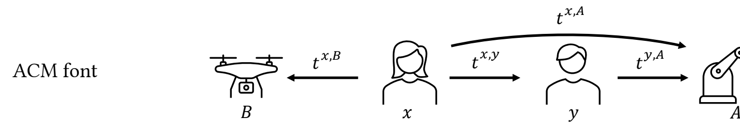
ACM Reference Format:

Yaohui Guo, X. Jessie Yang, and Cong Shi. 2023. TIP: A *T*rust *I*nference and *P*ropagation Model in Multi-Human Multi-Robot Teams. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580164

1 INTRODUCTION

Trust has been identified as one central factor for effective human-robot teaming [14, 17, 18]. Despite research efforts over the past thirty years, existing literature predominantly focuses on dyadic human-robot teams where one human agent interacts with one robot [12]. There is little, if not no, research on trust modeling in teams consisting of multiple human agents and multiple robots.

Consider a scenario where two human agents, x and y, and two robots, A and B, are to perform a task. The four agents are allowed to form sub-teams to enhance task performance (e.g., maximizing throughput and/or minimizing task completion time). For instance,



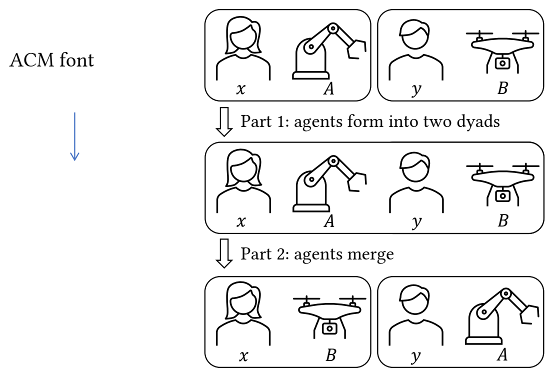


Figure 1: Four agents can form sub-teams. In Part 1, human x and robot A form a dyad, and human y and robot B form a dyad. In part 2, two dyads merge. In part 3, human x and robot B form a dyad, and human y and robot A form a dyad.

In this scenario, we assert that there exist two types of experiences that a human agent has with a robot: *direct and indirect experiences*. Direct experience, by its name, means that a human agent's interaction with a robot is by him-/her-self; indirect experience means that a human agent's interaction with a robot is mediated by another party. Fig. 2 illustrates Part *3* of the task shown in Fig. 1. Human x works directly with robot B (i.e., direct experience). Even though there is no direct interaction between x and A in part 3, we postulate that x could still update his or her trust in A by learning y's experience with A, i.e., y's direct experience with A becomes x's indirect experience. y's trust in A *propagates* to x.

Under the direct and indirect experience framework, prior work on trust modeling in dyadic human-robot teams can be regarded as examining how *direct* experience influences a person's trust in a robot. In multi-human multi-robot teams, we postulate that *both direct and indirect experiences drive a human agent's trust in a robot.*

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03…$15.00
https://doi.org/10.1145/3568294.3580164

CM font

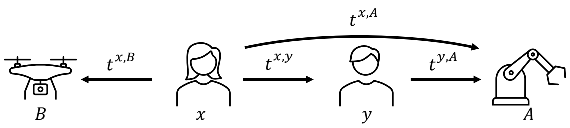
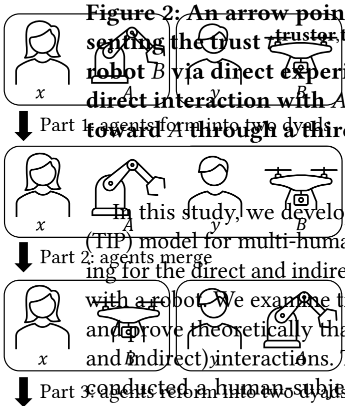


Figure 2: An arrow points from a trustor to a trustee, representing the trust $t^{\text{trustor,trustee}}$. Human x updates her trust in robot B via direct experience. Even though x does not have direct interaction with A, x could still update his or her trust toward A through a third party, y.



In this study, we develop the **T**rust **I**nference and **P**ropagation (TIP) model for multi-human multi-robot teams, explicitly accounting for the direct and indirect experiences a human agent may have with a robot. We examine trust dynamics under the TIP framework and prove theoretically that trust converges after repeated (direct and indirect) interactions. To evaluate the proposed TIP model, we conducted a human-subject experiment with 15 pairs of participants ($N = 30$). Each pair worked with two drones to perform a threat detection task for 15 sessions. We compared the TIP model (i.e., accounts for both the direct and indirect experiences) and a direct-experience-only model (i.e., only accounts for the direct experience a human agent has with a robot). Results show that the TIP model successfully captured people's trust dynamics with a significantly smaller root-mean-square error (RMSE) compared to the direct-experience-only model. To the best of our knowledge, the proposed TIP model is the first mathematical framework for computational trust modeling in multi-human multi-robot teams.

2 RELATED WORK

Several computational trust models in dyadic human-robot teams exist [3, 6, 13, 16]. Notably, Xu and Dudek [16] proposed the online probabilistic trust inference model (OPTIMo) utilizing Bayesian networks to estimate human trust from automation's performance and human behavior. Hu et al. [8] proposed to classify trust or distrust based on electroencephalography (EEG) and galvanic skin response (GSR) signals. Soh et al. [15] modeled trust as a latent dynamic function for predicting human trust in robots across different tasks. Guo and Yang [6] and Bhat et al. [1] proposed to model trust as a Beta random variable and predicted trust value in a Bayesian framework. For a detailed review, refer to [11].

Even though the amount of research on trust modeling in multi-human multi-robot teams is extremely limited, some inspiration can be drawn from research on reputation/trust management. Central to the reputation management system is a propagation mechanism that allows a buyer to obtain the reputation/trustworthiness values of a seller, especially when the buyer had no prior transactions with the seller [7]. Examples include the Beta reputation system [10] and the FIRE trust management model [9]. Variants of such propagation mechanisms can be found in several multi-agent systems, such as e-commerce [2] and social networks [4].

3 MATHEMATICAL MODEL

We present the detailed TIP model. Our objective is to develop a fully computational trust propagation model that works in general human-robot interaction settings.

3.1 Assumptions

We make two major assumptions. First, we assume that each human agent communicates trust as a single-dimensional value [5, 6]. Second, we assume that human agents are cooperative, i.e., they report their trust truthfully to their human teammates.

3.2 Proposed Model

Trust as a Beta random variable. We define a trust value as a real number in $[0, 1]$, where 0 stands for "[do] not trust at all" and 1 stands for "trust completely". We take a probabilistic view to model trust as in [6]. At time k, the trust $t_k^{a,b}$ that a human agent a has towards another agent b follows a Beta distribution, i.e., $t_k^{a,b} \sim \text{Beta}\left(\alpha_k^{a,b}, \beta_k^{a,b}\right)$, where $\alpha_k^{a,b}$ and $\beta_k^{a,b}$ are the positive and negative experiences a had about b up to time k, respectively, $k = 0, 1, 2, \dots$. When $k = 0$, $\alpha_0^{a,b}$ and $\beta_0^{a,b}$ are the prior experiences that a had before any interaction with b. The expected trust is given by $\mu_k^{a,b} = \alpha_k^{a,b} / \left(\alpha_k^{a,b} + \beta_k^{a,b}\right)$. Here we note that $t_k^{a,b}$ is the self-reported trust given by the agent a, which has some randomness due to subjectivity, while $\mu_k^{a,b}$ is the expected trust determined by the experiences.

Trust update through direct experience. Similar to [6], we update the direct trust experience at time k by setting

$$\begin{aligned} \alpha_k^{a,b} &= \alpha_{k-1}^{a,b} + s^{a,b} \cdot p_k^b \\ \beta_k^{a,b} &= \beta_{k-1}^{a,b} + f^{a,b} \cdot \overline{p}_k^b \end{aligned}. \tag{1}$$

Here p_k^b and $\overline{p}_k^b$ are the measurements of b's success and failure during time k, respectively; $s^{a,b}$ and $f^{a,b}$ are a's unit experience gains with respect to success or failure of b, respectively. We require $s^{a,b}$ and $f^{a,b}$ to be positive to ensure that cumulative experiences are non-decreasing. The updated trust $t_k^{a,b}$ follows the distribution $\text{Beta}(\alpha_k^{a,b}, \beta_k^{a,b})$.

Trust update through indirect experience propagation. Let x and y denote two human agents and let A denote a robot agent, as illustrated in Fig. 2. At time k, y communicates his or her trust $t_k^{y,A}$ on A with x, and then x updates his or her indirect trust experience by

$$\begin{aligned} \alpha_k^{x,A} &= \alpha_{k-1}^{x,A} + \hat{s}^{x,A} \cdot t_k^{x,y} \cdot \left[t_k^{y,A} - t_{k-1}^{x,A} \right]^+ \\ \beta_k^{x,A} &= \beta_{k-1}^{x,A} + \hat{f}^{x,A} \cdot t_k^{x,y} \cdot \left[t_{k-1}^{x,A} - t_k^{y,A} \right]^+ \end{aligned}, \tag{2}$$

where the superscript '+' means taking the positive part of the corresponding number, i.e., $t^+ = \max\{0, t\}$ for a real number t, and $t_k^{x,A} \sim \text{Beta}(\alpha_k^{x,A}, \beta_k^{x,A})$.

The intuition behind this model is that x needs to reason upon $t_k^{y,A}$, i.e., y's trust towards A. First, x compares y's trust $t_k^{y,A}$ with his or her previous trust $t_{k-1}^{x,A}$. Let $\Delta t := t_k^{y,A} - t_{k-1}^{x,A}$ be the difference. If $\Delta t \geq 0$, x gains positive indirect experience about A, which amounts to the product of the trust difference Δt, a coefficient $\hat{s}^{x,A}$, and a discounting factor $t_k^{x,y}$, i.e., x's trust on y; if $\Delta t < 0$, then x gains negative indirect experience about A, which is defined similarly.

3.3 Parameter Inference

The proposed model characterizes a human agent's tr bot by six parameters. For instance, the parameter of A, which is defined as $\theta^{x,A} = \left(\alpha_0^{x,A}, \beta_0^{x,A}, s^{x,A}, f^{x,A},\right.$ including x's prior experiences $\alpha_0^{x,A}$ and $\beta_0^{x,A}$, the uni perience gains s_A^x and f_A^x, and the unit indirect experi $\hat{s}_A^x$ and $\hat{f}_A^x$. We denote the indices of x's direct and in updates with A up to time k as D_k and $\overline{D}_k$, respectively can compute $\alpha_k^{x,A}$ and $\beta_k^{x,A}$, according to Eqs. (1) and (

$$\alpha_k^{x,A} = \alpha_0^{x,A} + s^{x,A} \sum_{j\in D_k} p_j^A + \hat{s} \sum_{j\in \overline{D}_k} t_j^{x,y} \left[t_j^{y,A} - t_{j-}^{x,}\right.$$

$$\beta_k^{x,A} = \beta_0^{x,A} + f^{x,A} \sum_{j\in D_k} \overline{p}_j^A + \hat{f} \sum_{j\in \overline{D}_k} t_j^{x,y} \left[t_{j-1}^{x,A} - t_j^{y,A} \right]$$

The optimal parameter $\theta_*^{x,A}$ maximizes the log likelihood function

$$H\left(\theta^{x,A}\right) := \sum_{k=0}^{K} \log \mathrm{Beta}\left(t_k^{x,A} \middle| \alpha_k^{x,A}, \beta_k^{x,A}\right), \quad (4)$$

where $\alpha_k^{x,A}$ and $\beta_k^{x,A}$ are defined in Eq. (3).

We note that $\log \mathrm{Beta}(t_k^{x,A}|\alpha_k^{x,A}, \beta_k^{x,A})$ is concave in $\theta^{x,A}$ by the composite rule that the function is concave in $(\alpha_k^{x,A}, \beta_k^{x,A})$ and $\alpha_k^{x,A}$ and $\beta_k^{x,A}$ are non-decreasing linear functions of $\theta^{x,A}$. Consequently, $H(\theta^{x,A})$ is concave in $\theta^{x,A}$ because it is a summation of several concave functions. Therefore, we can run the gradient descent method to compute the optimal parameters.

4 HUMAN-SUBJECT STUDY

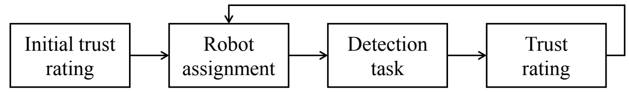
We conducted a human-subject experiment to evaluate the proposed model. The experiment, inspired by [19], simulated a threat detection task, where two human agents work with two smart drones to search for threats at multiple sites.

4.1 Participants

A total of $N = 30$ participants (average age = 25.3 years, SD = 4.3 ages, 16 females, 14 males) with normal or corrected-to-normal vision formed 15 teams and participated in the experiment. Each participant received a base payment of \$15 and a bonus of up to \$10 depending on their team performance. To promote cooperation between a pair of players, team performance instead of individual performance was used to calculate the bonus.

4.2 Experimental Task and Design

In the experiment, a pair of participants performed a simulated threat detection task with two assistant drones for $K = 15$ sessions on two separate desktop computers. At each session, each participant was assigned one drone and worked on the detection tasks. After the session, they were asked to report their trust in each drone and their trust in their human teammate. For clarity, we named the two drones A and B and colored them in red and blue, respectively; and we denoted the participants x and y. A trust rating is denoted as $t_k^{a,b}$, where the superscript $a \in \{x, y\}$ stands for the trustor, the superscript $b \in \{x, y, A, B\}$ stands for the trustee, and the subscript



(a) Flow of experimental task

(b) Task interface. The drone will highlight the potential threat in bright red. The participant is asked to click the 'Danger' button if a threat is present and to click the 'Clear' button otherwise.

Figure 3: Experimental task and design

k is the session index. For example, $t_2^{x,A}$ is person x's trust in drone A after the 2nd session. The range of a trust rating is $[0, 1]$, where 0 stands for "(do) not trust at all" and 1 stands for "trust completely". The flow of the experimental task is illustrated in Fig. 3a.

Initial trust rating: At the start, each participant gave their initial trust in the two drones based on their prior experience with automation/robots. Additionally, they gave their initial trust in each other. These trust ratings were indexed by 0, e.g., x's initial trust rating on A was denoted as $t_0^{x,A}$.

Robot assignment: At each session, each participant was randomly assigned one drone as his or her assistant robot.

Detection task: Each session consisted of 10 locations to detect. As shown in Fig. 3b, four views were present at each location. If a threat, which appeared like a combatant, was in any of the views, the participant should click the 'Danger' button; otherwise, they should click the 'Clear' button. Meanwhile, his or her drone would assist and highlight a view if the drone detected a threat there. In addition, a 3-second timer was set for each location. If a participant did not click either button before the timer counted down to zero, the testbed would move to the next location automatically. After all the 10 locations, an end-of-session screen was shown, displaying how many correct choices the participant and the drone had made in the current session. Correct choices mean correctly identifying threats or declaring 'Clear' within 3 seconds.

Trust rating: After each session, participants reported three trust values. First, each participant updated his or her trust in the drone s/he just worked with, i.e., through direct experience. Next, each participant submitted and communicated their trust score to their human teammate. After that, each participant updated his or her trust in the drone the human teammate just worked with (i.e., the other drone) and his or her trust in the human teammate. After participants completed all 15 sessions, the experiment ended.

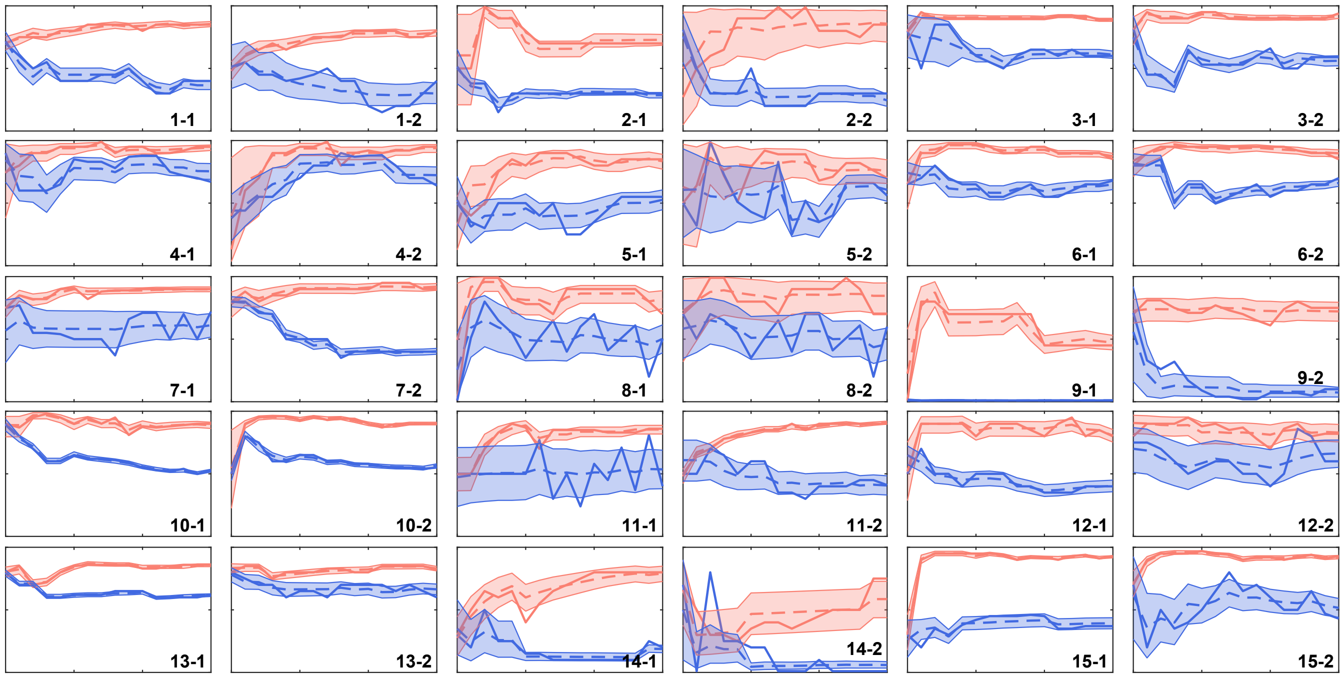


Figure 4: Fitting results. Red curves are for drone A while blue curves are for drone B. The solid lines are the participants' self-reported trust, while the dashed lines are the expected trust value predicted by the model. The shaded areas indicate the 90% probability interval of the Beta distribution at each session. The index i-j stands for the jth participant in the $i$the group.

4.3 Experimental Procedure

Before the experiment, each participant signed a consent form and filled out a demographic survey. Two practice sessions were provided, wherein a practice drone was used to assist the participants. The participants were told that the practice drone differed from the two drones used in the real experiment. After the experiment started, the assignment of drones was randomized in each group. Specifically, we assigned drone A with equal change to either participant and then assigned drone B to the other participant. The threat detection accuracy of the practice drone, drone A, and drone B were set to 80%, 90%, and 60%, respectively.

5 RESULTS AND DISCUSSION

We use the gradient descent method in Sec. 3.3 to compute the optimal parameters $\theta_*^{p_i,A}$ and $\theta_*^{p_i,B}$ for each participant p_i. The fitting results are shown in Fig. 4. We calculate the performance measurements of drone A at session k as $p_k^A = A_k/10$ and $\overline{p}_k^A = 1 - p_k^A$, where A_k is the number of correct choices drone A made in the kth session; and we define p_k^B and $\overline{p}_k^B$ similarly. To measure the performance of the model, we calculate the fitting error at each session for each participant as $e_k^{p_i,R} = |\mu_k^{p_i,R} - t_k^{p_i,R}|$, $R \in \{A, B\}$, where $t_k^{p_i,R}$ is the participant's self-reported trust while $\mu_k^{p_i,R}$ is the expected trust defined in section 3.2 and computed based on $\theta_*^{p_i,R}$; and, we calculate the root-mean-square error (RMSE) between the ground truth and the expected trust value as $\text{RMSE}^R = [\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K+1}\sum_{k=0}^{K}(e_k^{p_i,R})^2]^{1/2}$, for $R \in \{A, B\}$. The RMSE results for the TIP model are $\text{RMSE}^A = 0.057$ and $\text{RMSE}^B = 0.082$.

Fig. 4 shows the fitting results of the TIP model. The shaded regions indicate the 90% confidence interval of the Beta distribution at each session. We observe that for most participants, such as 7-2 and 10-2, the proposed TIP model can accurately fit the trust curve with a narrow confidence interval; but for some other participants, such as 5-2 and 8-1, the model cannot fit the trust curve due to trust oscillation. However, in the latter case, the fitted curve has a similar trend with the ground truth and can cover most data points with the 90% confidence interval.

For comparison, we consider a direct-update-only model that only accounts for the direct experience a human agent has with a robot. The direct-update-only model is equivalent to the TIP model with zero unit indirect experience gains, i.e., $\hat{s}^{x,A} = \hat{f}^{x,A} = 0$. We recompute the model parameters for the direct-update-only model, and the corresponding RMSE errors are $\text{RMSE}'^A = 0.085$ and $\text{RMSE}'^B = 0.107$. Furthermore, we compare each participant's fitting error $\bar{e}^{p_i,R} := 1/(K+1)\sum_{k=0}^{K} e_k^{p_i,R}$ of the TIP model (A: 0.044 ± 0.037; B: 0.069 ± 0.045) and that of the direct-update-only model (A: 0.075 ± 0.041; B: 0.095 ± 0.051) using a paired-sample t-test. Results show that the former is significantly smaller than the latter, with $t(29) = -6.18, p < .001$ for drone A, and $t(29) = -7.31$, $p < .001$ for drone B.

6 ACKNOWLEDGEMENT

This work is supported by the National Science Foundation under Grant No. 2045009 and the Air Force Office of Scientific Research under Grant No. FA9550-20-1-0406.

REFERENCES

[1] Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. 2022. Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8815–8822. https://doi.org/10.1109/LRA.2022.3188902

[2] Yukuo Cen, Jing Zhang, Gaofei Wang, Yujie Qian, Chuizheng Meng, Zonghong Dai, Hongxia Yang, and Jie Tang. 2020. Trust Relationship Prediction in Alibaba E-Commerce Platform. *IEEE Transactions on Knowledge and Data Engineering* 32, 5 (2020), 1024–1035. https://doi.org/10.1109/TKDE.2019.2893939

[3] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with Trust for Human-Robot Collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 307–315. https://doi.org/10.1145/3171221.3171264

[4] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph Neural Networks for Social Recommendation. In *The World Wide Web Conference on - WWW '19*. ACM Press, San Francisco, CA, USA, 417–426. https://doi.org/10.1145/3308558.3313488

[5] Yaohui Guo, Cong Shi, and X. Jessie Yang. 2021. Reverse psychology in trust-aware human-robot interaction. *IEEE Robotics and Automation Letters* 6, 3 (2021), 4851–4858. https://doi.org/10.1109/LRA.2021.3067626

[6] Yaohui Guo and X. Jessie Yang. 2021. Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics* 13 (2021), 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

[7] Ferry Hendrikx, Kris Bubendorfer, and Ryan Chard. 2015. Reputation systems: A survey and taxonomy. *J. Parallel Distrib. Comput.* 75 (2015), 184–197. https://doi.org/10.1016/j.jpdc.2014.08.004

[8] Wan Lin Hu, Kumar Akash, Neera Jain, and Tahira Reid. 2016. Real-Time Sensing of Trust in Human-Machine Interactions. *IFAC-PapersOnLine* 49, 32 (Jan. 2016), 48–53. https://doi.org/10.1016/j.ifacol.2016.12.188

[9] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. 2006. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 13, 2 (2006), 119–154. https://doi.org/10.1007/s10458-005-6825-4

[10] Audun Josang and Roslan Ismail. 2002. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, Vol. 5. 2502–2511.

[11] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1, 4 (2020), 297–309. https://doi.org/10.1007/S43154-020-00029-Y

[12] National Academies of Sciences, Engineering, and Medicine. 2022. *Human-AI Teaming: State-of-the-Art and Research Needs*. The National Academies Press, Washington, DC. https://doi.org/10.17226/26355

[13] Charles Pippin and Henrik Christensen. 2014. Trust modeling in multi-robot patrolling. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 59–66. https://doi.org/10.1109/ICRA.2014.6906590

[14] Thomas B. Sheridan. 2016. Human–Robot Interaction: Status and Challenges. *Human Factors* 58, 4 (2016), 525–532. https://doi.org/10.1177/0018720816644364

[15] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. 2020. Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research* 39, 2-3 (2020), 233–249. https://doi.org/10.1177/0278364919866905

[16] Anqi Xu and Gregory Dudek. 2015. OPTIMo. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. ACM Press, New York, New York, USA, 221–228. https://doi.org/10.1145/2696454.2696492

[17] X. Jessie Yang, Yaohui Guo, and Christoper Schemanske. 2023. From Trust to Trust Dynamics: Combining Empirical and Computational Approaches to Model and Predict Trust Dynamics In Human-Autonomy Interaction. In *Human-Automation Interaction: Transportation*, Vincent G. Duffy, Steven J. Landry, John D. Lee, and Neville Stanton (Eds.). Springer International Publishing, Cham, 253–265. https://doi.org/10.1007/978-3-031-10784-9_15

[18] X. Jessie Yang, Christopher Schemanske, and Christine Searle. 2021. Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation. *Human Factors* (2021), 00187208211034716. https://doi.org/10.1177/00187208211034716

[19] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*. ACM, New York, NY, USA, 408–416. https://doi.org/10.1145/2909824.3020230