A Unified Bi-directional Model for Natural and Artificial Trust in Human–Robot Collaboration

Hebert Azevedo-Sa[1], X. Jessie Yang[1,2], Lionel P. Robert Jr.[1,3] and Dawn M. Tilbury[1,4]

***Abstract*—We introduce a novel capabilities-based bi-directional multi-task trust model that can be used for trust prediction from either a human or a robotic trustor agent. Tasks are represented in terms of their capability requirements, while trustee agents are characterized by their individual capabilities. Trustee agents' capabilities are not deterministic; they are represented by belief distributions. For each task to be executed, a higher level of trust is assigned to trustee agents who have demonstrated that their capabilities exceed the task's requirements. We report results of an online experiment with 284 participants, revealing that our model outperforms existing models for multi-task trust prediction from a human trustor. We also present simulations of the model for determining trust from a robotic trustor. Our model is useful for control authority allocation applications that involve human–robot teams.**

***Index Terms*—Acceptability and Trust; Human-Robot Collaboration; Social HRI.**

I. Introduction

WOULD you *trust* someone to drive you in an overcrowded city with heavy traffic? You probably would, if you knew that person was a capable driver. Most certainly, to ultimately gain your trust, the potential *trustee* driver must demonstrate her/his competence by providing you—the *trustor* passenger—with a positive experience.

The driver–passenger example is only one of countless situations involving trust between two agents in a trust relationship: the trustor (the one who trusts) and the trustee (the one to be trusted). Trust pervades our relationship with other people, with organizations, and with machines [1]–[4]. Trust depends on both the trustor's and the trustee's characteristics and is revealed when the trustor takes the risk of being vulnerable to the trustee's actions [2].

Human–robot interaction (HRI) researchers have proposed predictive trust models that try to capture how a human trustor develops trust in a robotic trustee [5]–[7]. A perspective that is generally overlooked, however, is how trust from a robotic trustor should develop over interactions with a trustee. In this work we distinguish between human trust, which we label as *natural trust*, from robotic trust, which we label as *artificial trust*. Current trust models are focused on natural trust and are useful for trust-aware decision-making, which requires the robot to estimate the human's trust in the robot to plan actions in an HRI setting.

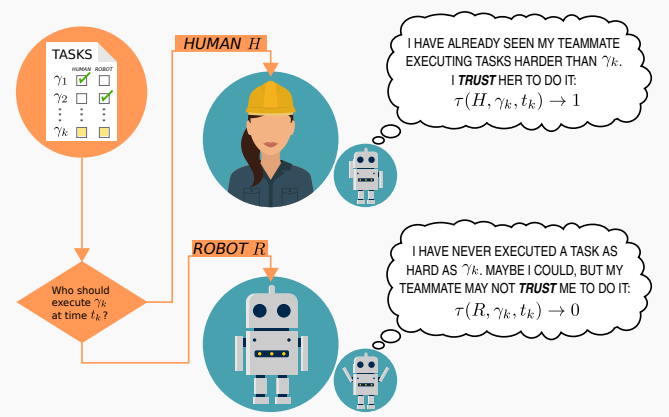

Fig. 1. A team formed by human H and a robot R that must collaborate sequentially executing tasks. Each task must be executed by one of the agents. The bi-directional trust model can be used for predicting a human's trust in a robot to execute a task, and to predict how much humans can be trusted to execute a task.

Existing trust models have several shortcomings that hinder their ability to predict humans' natural trust and limit their application for robots' artificial trust computation. First, current trust models are limited in their ability to characterize the tasks that should be executed by trustees. Tasks must be characterized in terms of what capabilities and which proficiency levels (in those capabilities) are required from trustees to execute those tasks. For instance, driving requires certain levels of cognitive, sensory and physical capabilities from drivers [8]. Second, current trust models fall short of describing the trustee agents in terms of their proven capabilities. Trustees' capabilities characterization and quantification are important because, when a trustor knows that the trustee is or is not capable of meeting the task requirements, the trustor's trust in the trustee to execute that task is higher (or lower). Finally, because of a lack of trustee capability characterization, current trust models are applicable for natural trust, or understanding human trust in a robot, but not for artificial trust, especially for determining how a

Manuscript received: February, 23, 2021; Revised May, 12, 2021; Accepted June, 2, 2021.

This paper was recommended for publication by Editor Gentiane Venture upon evaluation of the Associate Editor and Reviewers' comments.

This work was partially supported by the National Science Foundation and by the Brazilian Army's Department of Science and Technology.

[1]Hebert Azevedo-Sa, X. Jessie Yang, Lionel P. Robert Jr. and Dawn M. Tilbury are with the Robotics Institute, University of Michigan, Ann Arbor, MI, 48109 USA. {azevedo, xijyang, lprobert, tilbury}@umich.edu.

[2]X. Jessie Yang is with the Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor.

[3]Lionel P. Robert Jr. is with the School of Information, University of Michigan, Ann Arbor.

[4]Dawn M. Tilbury is with the Department of Mechanical Engineering, University of Michigan, Ann Arbor.

Digital Object Identifier (DOI): see top of this page.

robot should trust a human. Existing models are performance-centric and ignore *non-performance trustees' capabilities* or *factors*, which are needed for determining artificial trust. To accommodate both natural and artificial trust in (human or robotic) trustees, a computational model of trust must be able to consider assessments of a trustee's non-performance capabilities, such as honesty, benevolence or integrity levels [2], [9]. Therefore, although existing trust models are sufficient for planning algorithms, these trust models can not be used in more sophisticated control authority allocation applications, which are likely to be based on comparisons between the human's trust in the robot and the robot's trust in the human [10].

To address those shortcomings, we propose a novel capabilities-based bi-directional trust model. Our model characterizes tasks on a set of standard requirements that can represent either performance or non-performance capabilities that affect trust, and builds trustee capability profiles based on the trustee's history on executing those tasks. Trust is represented by the probability that an agent can successfully execute a task, considering that agent's capability profile (built after observations). By considering the agent's capabilities (performance or non-performance) [9] and the task requirements, our model can be used to determine a robot's artificial trust in a trustee agent. Moreover, our model can be used for predicting trust transfer between tasks, similar to the model proposed in [6]. However, as compared to [6], our model improves trust transfer predictions by representing tasks in terms of capability requirements instead of using natural language processing (NLP) similarity metrics. We show the superiority of our trust model by comparing its prediction results with those from other models, using a dataset collected in an online experiment with 284 participants. In sum, our contributions with this work are:

- a new trust model that (i) can be used for the *artificial* trust computation and (ii) outperforms existing models for multi-task *natural* trust transfer prediction; and
- an online experiment that resulted in a dataset relating trust and task capabilities measurements.

II. Trust in Human–Robot Interaction

A. Origins and Current Stage of Trust in HRI

Trust in robots that interact with humans can be considered as an evolution of trust in automation, which in turn has evolved from theoretical frameworks on interpersonal trust. Muir [11] proposed the concept of trust in automation after adapting sociologist interpersonal trust definitions [1], [12] to humans and automated machines [13]. Trust in automation is a dynamic construct [14] that can be directly measured with subjective scales [3], [15] or can also be estimated through behavioral variables [16], [17].

People's trust in an automated system must be calibrated, which means it has to align with the system's capabilities. Miscalibrated trust is likely to lead to the inappropriate use of the system [14], [18]–[20]. However, the evolution of automated systems into autonomous robots with powerful sensing technologies has paved the way for new trust calibration strategies. Robots can now perceive and process humans' trust and take action to increase or decrease humans' trust when necessary [20]–[22].

B. Trust Definition

Several trust definitions have been proposed, generally pointing to the trustor's attitude or willingness to be vulnerable to the trustee's actions [2], [4]. In this work, we assume the (adapted) definition for trust recently proposed by Kok and Soh, which states that: "given a trustor agent A and a trustee agent B, A's trust in B is a multidimensional latent variable that mediates the relationship between events in the past and A's subsequent choice of relying on B in an uncertain environment" [19]. Kok and Soh's definition establishes important aspects of our model, such as the multidimensionality of trust and its dependence on a history of events involving the trustor and the trustee agents.

C. Trust Computational Models

Trust models are usually applied to determine how much a human trusts a robot to perform a task (e.g. Fig. 1, where the robot R is chosen to execute a task). The robot uses this estimate of human trust to predict the human's behavior, such as intervening on the task execution. For example, trust models have been used in different trust-aware POMDP-based algorithms proposed for robotic planning and decision-making [22], [23]. Their objective is to eventually improve the robot's collaboration with the human, using human trust as a vital factor when planning the robot's actions.

Planning and decision-making frameworks usually rely on the use of probabilistic models for trust [5], [24], [25]. Xu and Dudek proposed an online probabilistic trust inference model for human–robot collaborations (OPTIMo) that uses a dynamic Bayesian network (DBN) combined with a linear Gaussian model and recursively reduces the uncertainty around the human operator's trust. OPTIMo was tested in a human–unmanned aerial vehicle (UAV) collaboration setting [5] and, although some dynamic models had been proposed before [13], [26], OPTIMo was the first trust model capable of tracking human's trust in a robot with low latency and relatively high accuracy. The UAV, with OPTIMo, was able to track the human operator's trust by observing how much the human intervened in the UAV's operation.

Other Bayesian models have been proposed since OPTIMo. These models include personalized trust models that apply inference over a history of robot performances, such as [25] and [24]. Mahani et al. proposed a model for trust in a swarm of UAVs, establishing a baseline for human–multi-robot interaction trust prediction [25]. Guo and Yang [24] have improved trust prediction accuracy (as compared to Lee's ARMAV model [13] and OPTIMo [5]) by proposing a formulation that describes trust in terms of Beta probability distributions and aligns the inference processes with trust formation and evolution processes [24]. Without explicitly modeling trust, Lee et al. showed that a robot that estimates and calibrates humans' intents and capabilities while making decisions can engender higher trust from humans [27].

Although all previously mentioned approaches for trust modeling represent important advances in how we understand and describe humans' trust in robots, they suffer from a common limitation. Those models depend on the history of robots' performances on unique specific tasks and are not applicable for trust transfer between different tasks. The issue of multi-task trust transfer was recently approached by Soh et al. [6], who proposed Gaussian processes and neural methods for predicting the transferred trust among different tasks that were described with NLP-based text embeddings. A major goal for our model was to deepen that discussion and improve prediction accuracy for multi-task trust transfer by (i) describing tasks in terms of capability requirements, and (ii) describing potential trustee agents in terms of their proven capabilities that can be used to transfer trust to another task.

The other major goal for our model was to be bi-directional, i.e., to be able to represent either natural trust or artificial trust. Because the existing trust models are usually performance-centric, they are suited to represent humans' natural trust in robots. Although mutual trust has been modeled as a single variable that depends on both the human's and the robot's performances on collaborative tasks [28], to represent a robot's artificial trust in humans, trust models must be more comprehensive. Computational models of trust must consider not only performance factors but also non-performance factors that describe human trustees [2], [9], [29], [30]. Until recently, only a few trust models have considered the robot's trust perspective, focusing only on non-performance factors that affect trust. For instance, a model that reproduces theory of mind (ToM) aspects in robots to identify deceptive humans has been proposed and applied in [29] and [30]. Our model is applicable for either natural or artificial trust because it explicitly considers a general form of agents' capabilities and task requirements, which can represent performance or non-performance trustee capabilities.

III. Bi-Directional Trust Model Development

A. Context Description

Consider the following situation: two agents (human H or robot R) collaborate and must execute a sequence of tasks. These tasks are indivisible and must be executed by only one agent. The execution of each task can either succeed or fail. For each task, one of the agents is in the position of trustor, and the other is the trustee. Therefore, the trustor is vulnerable to the trustee's performance in that task. From previous experiences with the trustee, the trustor has some implicit knowledge about the trustee's capabilities. This implicit knowledge is used by the trustor to assess how likely the trustee is to succeed or fail in the execution of a task. We define the terms and concepts we need for developing our trust model:

Definition 1 - Task. A *task* that must be executed is represented by $\gamma \in \Gamma$. Γ represents the set of all tasks that can be executed by the agents.

Definition 2 - Agent. An *agent* $a \in \{H, R\}$ represents a trustee that could execute a task γ.

Definition 3 - Capability. The representation of a specific skill that agents have/that is required for the execution of tasks. We represent a capability as an element of a closed interval $\Lambda_i = [0, 1]$, $i \in \{1, 2, 3, ..., n\}$, with n being a finite number of dimensions characterizing distinct capabilities.

Definition 4 - Capability Hypercube. The compact set representation of n distinct capabilities, given by the Cartesian product $\Lambda = \prod_{i=1}^{n} \Lambda_i = [0, 1]^n$. This definition is inspired by the particular capabilities from Mayer et al.'s model [2], namely ability, benevolence and integrity, but the definition is intended to be broader than these three dimensions.

Definition 5 - Agent's Capability Transform. The agent capability transform $\xi : \{H, R\} \rightarrow \Lambda$ maps an agent into a point in the capability hypercube representing the *agent's capabilities*, given by $\xi(a) = \lambda = (\lambda_1, \lambda_2, ..., \lambda_n) \in \Lambda$.

Definition 6 - Task Requirements Transform. The *task requirements transform* $\varrho : \Gamma \rightarrow \Lambda$ maps a task γ into the minimum required capabilities for the execution of γ, given by $\varrho(\gamma) = \bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2, ..., \bar{\lambda}_n) \in \Lambda$.

Definition 7 - Time Index. The *time* is discrete and represented by $t \in \mathbb{N}$.

Definition 8 - Task Outcome. The outcome of a task γ after being executed by the agent a at the time t is represented by $\Omega(\xi(a), \varrho(\gamma), t) \in \{0, 1\}$, where 0 represents a failure and 1 represents a success. We also define the Boolean complement of Ω, denoted by $\mho$, such that $\mho = 1$ when $\Omega = 0$, and $\mho = 0$ when $\Omega = 1$.

Leveraging the previous definitions, we can finally define trust.

Definition 9 - Trust. A trustor agent's trust in a trustee agent a to execute a task γ at a time instance t can be represented by

$$\begin{aligned} \tau(a, \gamma, t) &= P\big(\Omega(\xi(a), \varrho(\gamma), t) = 1\big) \\ &= \int_{\Lambda} p\big(\Omega(\lambda, \bar{\lambda}, t) = 1 | \lambda, t\big) bel(\lambda, t-1) d\lambda, \end{aligned} \tag{1}$$

where $\lambda = \xi(a)$, $\bar{\lambda} = \varrho(\gamma)$, and $bel(\lambda, t-1)$ represents the trustor's belief in the agent's capabilities λ at time $t-1$ (i.e., before the actual task execution). The belief is a dynamic probability distribution over the capability hypercube Λ. Note that, at each time instance t, trust is a function of the task requirements $\bar{\lambda}$, representing a *probability of success* in $[0, 1]$.

B. Bi-directional Trust Model

Our bi-directional model is defined by Eq. (1) and depends on the combination of:

- a function to represent the "trust given the trustee's capability", represented by the conditional probability $p\big(\Omega(\lambda, \bar{\lambda}, t) = 1 | \lambda, t\big)$; and
- a process to dynamically update the trustor's belief over the trustee capabilities $bel(\lambda, t)$.

We assume that an agent that successfully performs a task is more likely to be successful on less demanding tasks. Conversely, an agent that fails on a task is more likely to fail on more demanding tasks. We adapt the sigmoid function to represent that logic, and for each capability dimension we can write

$$\tau_i = \left[\frac{1}{1 + e^{\beta_i (\bar{\lambda}_i - \lambda_i)}} \right]^{\zeta_i}, \tag{2}$$

where $\beta_i, \zeta_i > 0$. Considering that all capability dimensions must be assessed concurrently and assuming that the capability dimensions are represented by independent random variables, for the probability computation, we have

$$p\big(\Omega(\lambda, \bar{\lambda}) = 1|\lambda\big) = \prod_{i=1}^{n} \tau_i = \prod_{i=1}^{n} \left[\frac{1}{1 + e^{\beta_i(\bar{\lambda}_i - \lambda_i)}}\right]^{\zeta_i}, \quad (3)$$

where t was suppressed, as the resulting function is independent of the time. The product of probabilities in Eq. (3) can quickly converge to zero as n increases. Therefore, to improve code implementation stability in practical implementations, a linear form of Eq. (3) could be used (i.e., by taking the logarithm on both sides of the equation).

Trust dynamics is established with a process for updating $bel(\lambda, t)$ that relates observations of a trustee agent's past performances with that agent's likelihood of success on related tasks. We considered that a trustor agent must build the belief about the trustee's capabilities after observations of the trustee's performances. However, initially, the trustor has no information about the trustee's performances and capabilities. We assumed this is represented by $bel(\lambda, 0)$ being a uniform probability distribution over the capability hypercube Λ, i.e., $bel(\lambda_i, 0) = \mathcal{U}(0, 1), \forall i \in \{1, 2, ..., n\}$. Next, after observing the sequence of successes and failures of the trustee in different tasks, the trustor updates $bel(\lambda, t)$, following the procedures in Algorithm 1 and in Fig. 2

Algorithm 1 Capability Belief Initialization and Update

procedure CAPABILITY HYPERCUBE INITIALIZATION
  for $i = 1 : n$ **do**
    $\ell_i \leftarrow 0$
    $u_i \leftarrow 1$
    $bel(\lambda_i, 0) \leftarrow \mathcal{U}(\ell_i, u_i)$ ▷ *Uniform distributions*
  end for
end procedure
procedure CAPABILITY UPDATE($\gamma, bel(\lambda, t-1)$)
  ▷ *When trustor observes trustee executing γ at t*
  for $i = 1 : n$ **do**
    if $\Omega(\lambda, \bar{\lambda}, t) = 1$ **then**
      if $\bar{\lambda}_i > u_i$ **then**
        $u_i \leftarrow \bar{\lambda}_i$
      else if $\bar{\lambda}_i > \ell_i$ **then**
        $\ell_i \leftarrow \bar{\lambda}_i$
      end if
    else if $\Omega(\lambda, \bar{\lambda}, t) = 0$ **then**
      if $\bar{\lambda}_i < \ell_i$ **then**
        $\ell_i \leftarrow \bar{\lambda}_i$
      else if $\bar{\lambda}_i < u_i$ **then**
        $u_i \leftarrow \bar{\lambda}_i$
      end if
    end if
    $bel(\lambda_i, t) \leftarrow \mathcal{U}(\ell_i, u_i)$
  end for
end procedure

C. Artificial Trust

For representing the artificial trust of a robotic trustor in a trustee agent, the bi-directional trust model can be slightly modified. We can vanish subjective biases that characterize human trustors by considering large values for the parameters β_i in Eq. (2) (i.e., considering the robot to be "infinitely pragmatic"). With sufficiently large β_i, τ_i becomes an analytic approximation of a decreasing step function with the transition from 1 to 0 when $\bar{\lambda}_i = \lambda_i$, i.e.

$$\lim_{\beta_i \to \infty} \tau_i = \mathcal{H}\big(-\bar{\lambda}_i + \lambda_i\big), \quad (4)$$

where $\mathcal{H}(x)$ is the Heaviside function of a continuous real variable x. Considering all capability dimensions to be independent, and using the approximation in Eq. (4) for computing trust with Eq. (3) and Eq. (1), we have

$$\tau(a, \gamma, t) = \prod_{i=1}^{n} \psi(\bar{\lambda}_i), \quad (5)$$

where,

$$\psi(\bar{\lambda}_i) = \begin{cases} 1 & \text{if } 0 \leq \bar{\lambda}_i \leq \ell_i, \\ \frac{u_i - \bar{\lambda}_i}{u_i - \ell_i} & \text{if } \ell_i < \bar{\lambda}_i < u_i, \\ 0 & \text{if } u_i \leq \bar{\lambda}_i \leq 1. \end{cases} \quad (6)$$

For each capability dimension, the robotic trustor agent believes that the trustee agent's capability is a random variable λ_i uniformly distributed between ℓ_i and u_i. If a task requires $\bar{\lambda}_i < \ell_i$, the trustee capability exceeds the task requirement and trust is 1. Conversely, if $\bar{\lambda}_i > u_i$, the task requirement exceeds the trustee's capability and trust is 0. In the intermediate condition, trust decreases with a constant slope from 1 to 0, corresponding to $\bar{\lambda}_i = \ell_i$ and $\bar{\lambda}_i = u_i$.

Robots can use long-term information to update their capability beliefs with a process different from that presented in Algorithm 1. An alternative is to recursively solve an optimization problem, considering the history of outcomes observed from different tasks γ (with different $\varrho(\gamma) = \bar{\lambda} \in \Lambda$). Trust is approximated by the number of successes divided by the number of times the task γ was performed, i.e.,

$$\hat{\tau} = \frac{\sum_{m=0}^{t} \Omega(\xi(a), \varrho(\gamma), m)}{\sum_{m=0}^{t} \big[\Omega(\xi(a), \varrho(\gamma), m) + \mho(\xi(a), \varrho(\gamma), m)\big]}, \quad (7)$$

and, considering each $\lambda = \varrho(\gamma)$, the capability distribution limits ℓ_i and u_i should be chosen such that $bel(\lambda, t) = \prod_{i=1}^{n} \mathcal{U}(\hat{\ell}_i, \hat{u}_i)$, and

$$(\hat{\ell}_i, \hat{u}_i) = \underset{[0,1]^2}{\arg\min} \int_{\Lambda} \|\tau - \hat{\tau}\|^2 d\lambda. \quad (8)$$

For numerical computations, Λ can be discretized and Eq. (8) approximated with a summation, as in Section V-B.

IV. EXPERIMENT

We conducted an online experiment using a Qualtrics survey and the Amazon Mechanical Turk (MTurk) platform to gather a dataset for comparing our model with other trust prediction

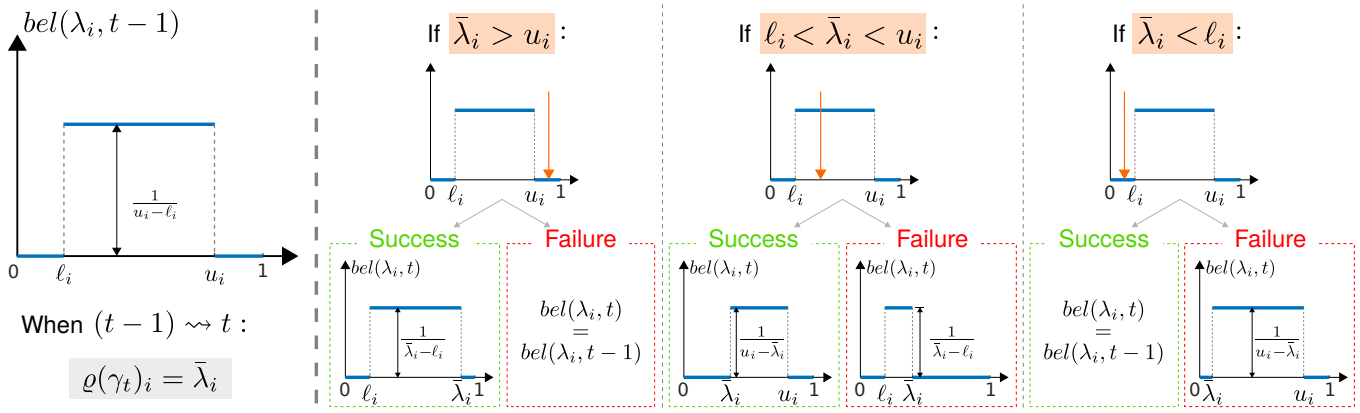

Fig. 2. Capability update procedure, where each capability dimension changes after the trustor agent observes the trustee agent a executing a task γ_t (at a specific time instance t). The belief distribution over a's capabilities *before* the task execution $bel(\lambda_i, t-1)$ is updated to $bel(\lambda_i, t)$, depending on the task capability requirements $\varrho(\gamma_t)_i = \bar{\lambda}_i$ and on the performance of a in γ_t, which can be a success ($\Omega = 1$) or a failure ($\Omega = 0$). The capability belief: (i) expands either when the agent succeeds on a task whose requirement exceeds u_i, or when the agent fails on a task whose requirement is less than ℓ_i; (ii) contracts when the agent succeeds or fails on a task whose requirement falls between u_i and ℓ_i; or (iii) remains the same either when the agent fails on a task whose requirement exceeds u_i, or when the agent succeeds on a task whose requirement is less than ℓ_i.

models, such as Soh's models [6] and OPTIMo [5]. We aimed to emulate a human-automated vehicle (AV) interaction setting, asking participants to (1) assess the requirement levels for driving tasks that were to be executed by the AV, (2) watch videos of the AV executing a part of those tasks and (3) evaluate their trust in the AV to execute other tasks (distinct from those they have watched in the videos).

Initially, only images and verbal descriptions of four driving tasks were presented in random order to the participants (Fig. 3). Participants were asked to rate the capability requirements for each of the presented tasks in terms of two distinct capabilities of the AV: sensing and processing, which were defined and presented to the participants as,

- **Sensing** (λ_s) **-** *The accuracy and precision of the sensors used to map the environment where the AV is located and perceive elements within that environment, such as other vehicles, people and traffic signs.*
- **Processing** (λ_p) **-** *The speed and performance of the AV's computers that use the information from sensors to calculate the trajectories and the steering, acceleration, and braking needed to execute those trajectories.*

Participants were asked to indicate the required capability levels $(\bar{\lambda}_s, \bar{\lambda}_p) \in [0,1]^2$ for each task, providing a score (i.e., indicating a slider position on a continuous scale) varying from low to high.

After evaluating all four presented tasks, participants watched short videos (approximately $20s$ to $30s$) of a simulated AV executing three of the four tasks. Those three were considered *observation tasks*. The videos showed the AV succeeding or failing to execute each observation task. (All videos are available at https://bit.ly/37gXXkI.) Next, participants were asked to indicate whether the AV successfully executed the task. That question served both as an attention checker and as a way to make the participant acknowledge the performance of the AV in that specific task. After watching each video, participants were also asked to rate their trust τ in the AV to execute the fourth remaining task (i.e., the *trust prediction task*) on a 7-point Likert scale varying from "very low trust" to "very high trust", as an indication of how much they disagreed or agreed with the sentence: "*I believe that the AV would successfully execute the task*." Participants were asked to consider all videos they had seen during the observation tasks and rate their trust in the AV to execute the trust prediction task. Finally, participants received a random 4-digit identifier code to upload in the MTurk platform and receive their payment.

To keep work-related regulations consistent, we restricted our participants to those who were physically in the United States when accepting the MTurk human intelligence task (HIT). A total of 284 MTurk workers participated in our experiment and received a payment of $1.80 for completing the HIT without failing to correctly answer the attention checker questions. The HITs were completed in approximately 6min40s, on average. We collected no demographics data or other personal information from the participants because these were not needed for our analyses. The obtained dataset and our implementations are available at https://bit.ly/3sfVtuK. The research was reviewed and approved by the University of Michigan's institutional review board (IRB# HUM00192470).

V. Results

A. Human-drivers' (natural) trust in robotic AVs

We implemented a 10-fold cross-validation to train and evaluate our bi-directional trust model (BTM) with the data obtained in the experiment described in Section IV. For comparison, we also evaluated the performance of Soh's Bayesian Gaussian process model (GP) [6] and that of a linear Gaussian model similar to Xu and Dudek's OPTIMo (OPT) [5] on our collected dataset. We obtained the tasks' vector representations for the GP model with GloVe [31], by processing the verbal descriptions presented in Fig. 3. There were no closed forms for Eq. (1), therefore we discretized each

Park, moving forward, in an empty space.

Park parallel to curb in a space between cars.

When reaching a roundabout, check left for oncoming traffic and complete the right turn when safe.

When navigating on a two-way road behind a vehicle and in foggy weather, check for oncoming traffic and pass when safe.

Fig. 3. Tasks presented to the experiment participants in terms of images and corresponding verbal descriptions. The participants had to rate the capability requirements for each of these tasks, considering two capability dimensions: sensing and processing. In other words, they had to assign a pair $(\bar{\lambda}_1, \bar{\lambda}_2) \in [0,1]^2$ for each task. Tasks were randomly presented for avoiding ordering effects.

task capability dimensions in 10 equal parts and computed numerical approximations for τ. Because we considered only two outcome possibilities (failure or success in executing a task), the trust measurements from both the dataset and the model outputs were considered probability parameters of Bernoulli distributions. We considered the cross entropy between those distributions to be the loss function to be minimized. We used PyTorch [32] to implement all parameter optimizations with the Adam algorithm [33], using randomized validation sets comprising 15% of the training data. Two metric scores were computed for the comparisons among model performances: the mean absolute error (MAE); and the negative log-likelihood (NLL), which corresponds to the loss function chosen for the optimizations.

Table I presents the MAE and NLL scores averaged over the 10 cross-validation folds (with standard deviations between parentheses) for the BTM, GP and OPT models. Fig. 4 complements the table, showing the average learning curves for both scores and bars representing the average final values with ± 1 standard deviations.

TABLE I
MEAN ABSOLUTE ERROR (MAE) AND NEGATIVE LOG-LIKELIHOOD (NLL) AVERAGE MINIMIZED SCORES FOR EACH TRUST MODEL

Model	MAE†	NLL†
BTM	$\mathbf{0.196(0.020)}^\ddagger$	$\mathbf{0.593(0.033)}^\ddagger$
GP	0.220(0.028)	0.619(0.060)
OPT	0.280(0.016)	0.672(0.021)

†10-fold results: Mean(Standard Deviation).
‡Best scores in **bold**.

Our bi-directional trust model (BTM) outperformed both the GP and the OPT models after the parameter optimization process. BTM reduced the MAE metric by approximately 11% as compared with GP, and by 30% as compared to OPT. In terms of NLL, the use of BTM reduced this metric by approximately 4.3% as compared with GP model, and by 12% as compared with the OPT model.

B. Robots' Artificial Trust in Humans

Besides evaluating and comparing our bi-directional trust model with other trust models using experimental data, we also implemented simulations to verify its use in the artificial trust mode (i.e., as a model for predicting a robots' trust in another trustee agent). We assumed two unspecified capability dimensions, considering that a trustee agent a's capabilities were static and represented by a point $\xi(a) = (\lambda_1, \lambda_2) \in \Lambda = [0,1]^2$. The trustee agent's capabilities were initially unknown by the trustor robot, who must estimate $\xi(a)$ after observing the trustee's performances in several different tasks. We considered N fictitious tasks γ^j, $j \in \{1, 2, ..., N\}$, and randomly picked N points $\varrho(\gamma^j) = (\bar{\lambda}_1^j, \bar{\lambda}_2^j) \in \Lambda$ representing capability requirements for the tasks. Task outcomes were assigned to each of the N tasks, with high probability of success for tasks that simultaneously had $\bar{\lambda}_1^j \leq \lambda_1$ and $\bar{\lambda}_2^j \leq \lambda_2$, and low probability of success when $\bar{\lambda}_1^j > \lambda_1$ or $\bar{\lambda}_2^j > \lambda_2$. Again, for numerical computations, we discretized both capability dimensions in 10 equal parts, obtaining 100 bins for Λ. We computed the observed probabilities of success

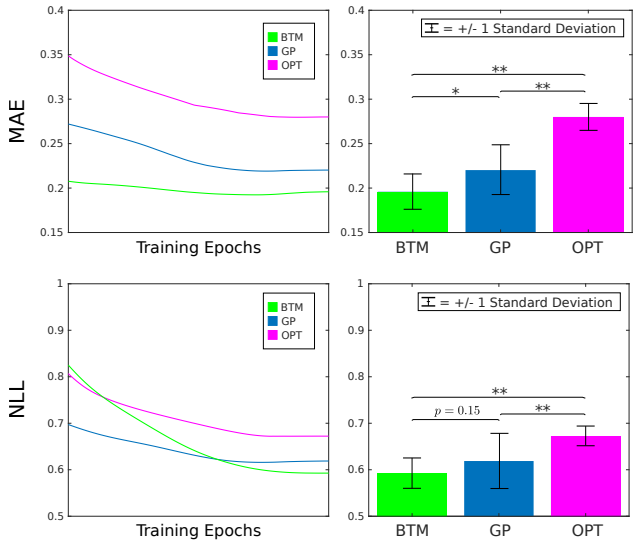

Fig. 4. MAE and NLL learning curves and final values for our proposed trust model (BTM) and for current trust models from [6] (GP) and [5] (OPT). As the total number of training epochs is different for each model, their representation on the horizontal axes of the learning curves is normalized.*$p < 0.05$; **$p < 0.01$.

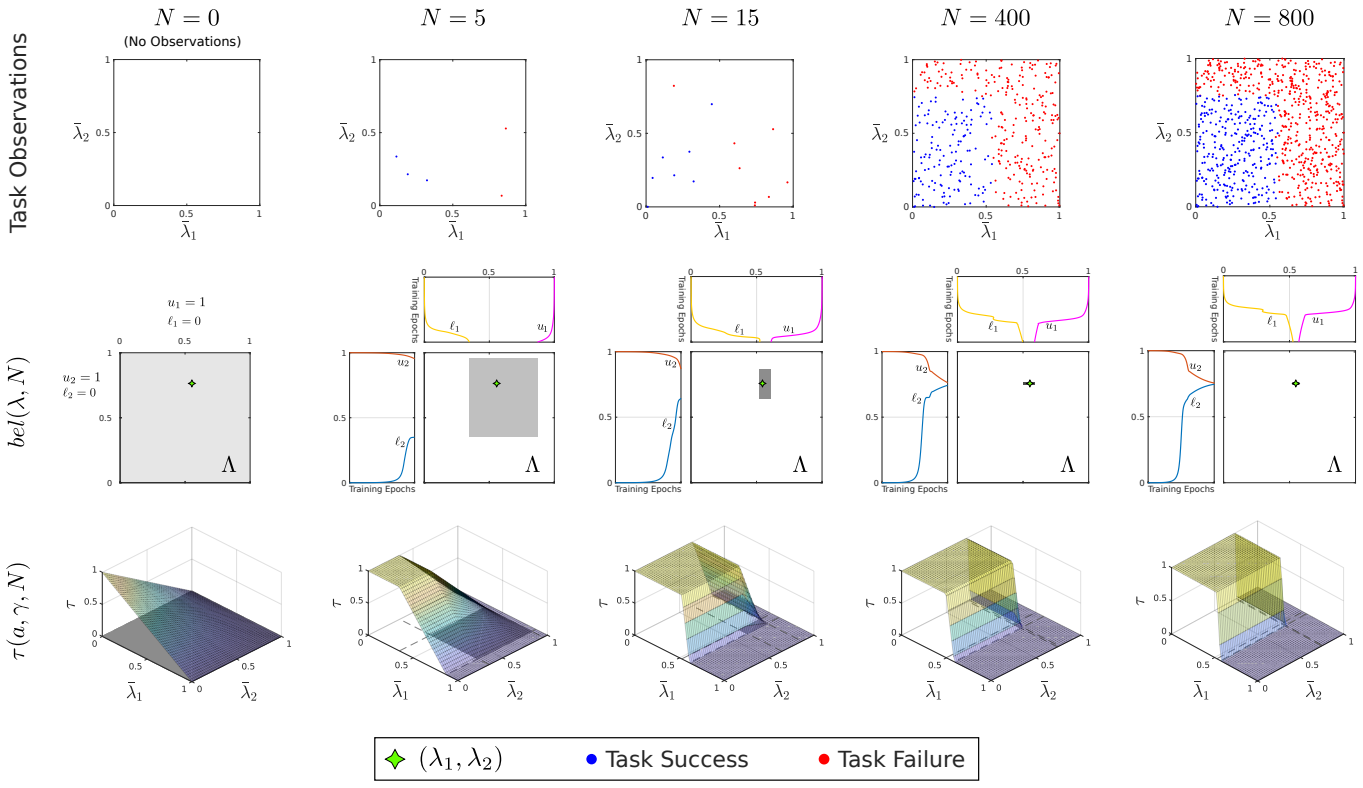

Fig. 5. Artificial trust results, where a robotic trustor agent's belief over a trustee agent a's capabilities is updated after N observations of a's performances in different tasks, represented by points in $\Lambda = [0, 1]^2$. When $N = 0$, $bel(\lambda, N)$ is "spread" over the entire Λ. When the robot trustor collects observations, it starts building a's capabilities profile and reducing the gray area in the $bel(\lambda, N)$ distribution. This profile gets more accurate when N increases and (λ_1, λ_2) gets better defined. This is also reflected in the evolution of the conditional trust function $\tau(a, \gamma, N)$.

for tasks inside a bin dividing the number of successes by the total number of tasks that fell on each bin (i.e., the approximation for $\hat{\tau}$). Finally we ran optimizations to find the parameters that best characterized $bel(\lambda_1, N)$ and $bel(\lambda_2, N)$, solving the problem represented by Eq. (8). Fig. 5 illustrates the evolution of $bel(\lambda, N)$ and of $\tau(a, \gamma, N)$ for increasing values of N. The higher the number of observations, the better the accuracy of a's identified capabilities.

VI. Discussion

Our model is based on general capability representations that can be either performance or non-performance trust factors. This particular aspect of our bi-directional trust model makes it useful for representing a robot's artificial trust, as presented in Subsection V-B, and allows for better human trust predictions in comparison to existing models, as presented in Subsection V-A. Additionally, our model considers task capability requirements in its description, describing how hard a task is for an agent to execute. The model's mathematical formulation captures the differences between those task requirements and the potential trustee agent's observed capabilities. Differently from the Gaussian process-based method presented in [6], this formulation allows for the adequate representation of lower trust levels when the requirements of a task exceed the capabilities of the agent and, conversely, higher trust levels when the agent capabilities exceed the task requirements.

The results reveal that our proposed bi-directional trust model has better performance for predicting a human's trust in a robot (in our specific experiment, an AV) than the models from [5] and [6]. This performance improvement was expected because current models are limited in capturing important trust-related parameters, such as the agents' capabilities or task's requirements in their formulation. To the best of our knowledge, only our model and Soh's models [6] distinguish and describe the trust transfer between different tasks, while OPTIMo [5] is more appropriate for predicting a human's trust in a robot to execute one specific task.

Section V-B presents simulations that show how the proposed model can be used for representing a robot's artificial trust. In the future, the proposed bi-directional trust model could be used in real-world human subjects experiments. An example could be a study where participants would execute some tasks represented in the capabilities hypercube, and the robot would be able to establish its trust in the participants based on their failures or successes on those tasks. In parallel, the robot could estimate the human's natural trust for different tasks, and use both natural and artificial trust metrics to compute expected rewards for the execution of new tasks. Tasks could be allocated between the human and the robot to maximize the expected reward of a whole set of tasks, eventually improving the joint performance of the human–

robot team.

Despite the eventual improvement on multi-task trust prediction performance, the use of task capability requirements could also be considered a drawback of our model because it calls for one more subjective input dimension in comparison with current models. Rating and describing tasks that must be executed by humans and robots in terms of specific human/robotic capability dimensions depends on the trustor agent's individual beliefs and experiences—natural, in the case of a human trustor agent, or artificial, in the case of a robotic trustor agent. Our models' trust prediction performance might have also been restricted by inconsistencies related to task characterization by each participant of our experiment. We believe that better trust prediction results can be achieved with in-person longitudinal experiments involving fewer participants and more predictions.

VII. Conclusion

We presented a multi-task bi-directional trust model that depends on both a trustee agent's proven capabilities (as observed by the trustor agent) and on the task capability requirements (as characterized by that same trustor agent). Our model outperformed the most relevant and recent trust models (i.e., [5] and [6]) in terms of predicting the transferred trust between distinct tasks by addressing the main limitations of those models. With a generalist capability dimension representing trustee agents' capabilities, our model can also represent robots' artificial trust in different trustee agents. Our model is useful for future applications where humans and robots collaborate and must sequentially take turns in executing different tasks.

References

[1] B. Barber, *The logic and limits of trust*. New Brunswick, NJ, USA: Rutgers Univ. Press, 1983, vol. 96.

[2] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," *Acad. Manage. Rev.*, vol. 20, no. 3, p. 709, Jul. 1995.

[3] B. M. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

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

[5] A. Xu and G. Dudek, "OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations," *in ACM/IEEE Int. Conf. on Human-Robot Interact.*, pp. 221–228, 2015.

[6] H. Soh, Y. Xie, M. Chen, and D. Hsu, "Multi-task trust transfer for human–robot interaction," *The Int. J. Robot. Res.*, vol. 39, no. 2-3, pp. 233–249, 2020.

[7] S. You and L. P. Robert, "Human-robot similarity and willingness to work with a robotic co-worker," in *Proc. 2018 ACM/IEEE Int. Conf. Human–Robot Interact.*, 2018.

[8] K. J. Anstey, J. Wood, S. Lord, and J. G. Walker, "Cognitive, sensory and physical factors enabling driving safety in older adults," *Clin. Psychol. Rev.*, vol. 25, no. 1, pp. 45–65, 2005.

[9] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interact.* Amsterdam, The Netherlands: Elsevier, 2021, pp. 3–25.

[10] H. Azevedo-Sa, X. J. Yang, L. Robert, and D. Tilbury, "Handling trust between drivers and automated vehicles for improved collaboration," in *2021 ACM/IEEE Int. Conf. Human–Robot Interact.* ACM, 2021.

[11] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man Mach. Stud.*, vol. 27, pp. 527–539, 1987.

[12] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships." *J. Pers. Soc. Psychol.*, vol. 49, no. 1, p. 95, 1985.

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

[14] M. Lewis, K. Sycara, and P. Walker, "The Role of Trust in Human-Robot Interaction," in *Studies in Systems, Decision and Control*. Berlin, Germany: Springer-Verlag, 2018, vol. 117, pp. 135–159.

[15] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergon.*, vol. 4, no. 1, pp. 53–71, 2000.

[16] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Hum-Comput. Stud.*, vol. 40, no. 1, pp. 153–184, 1994.

[17] H. Azevedo-Sa, S. K. Jayaraman, C. T. Esterwood, X. J. Yang, L. P. Robert, and D. M. Tilbury, "Real-Time Estimation of Drivers' Trust in Automated Driving Systems," *Int. J. Soc. Robot.*, pp. 1–17, 2020.

[18] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[19] B. C. Kok and H. Soh, "Trust in Robots: Challenges and Opportunities," *Curr. Robot. Rep.*, vol. 1, pp. 297–309, 2020.

[20] H. Azevedo-Sa, S. K. Jayaraman, X. J. Yang, L. P. Robert, and D. M. Tilbury, "Context-Adaptive Management of Drivers' Trust in Automated Vehicles," *IEEE Robot. Autom. Let.*, vol. 5, no. 4, pp. 6908–6915, 2020.

[21] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Trust-Aware Decision Making for Human-Robot Collaboration," *ACM Trans. Human-Robot Interact.*, vol. 9, no. 2, pp. 1–23, 2020.

[22] S. Sheng, E. Pakdamanian, K. Han, Z. Wang, J. Lenneman, and L. Feng, "Trust-based route planning for automated vehicles," in *12th ACM/IEEE Int. Conf. Cyber-Physic. Syst. (ICCPS '21)*. ACM, 2021.

[23] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with trust for human-robot collaboration," in *Proc. 2018 ACM/IEEE Int. Conf. on Human-Robot Interact.*, 2018, pp. 307–315.

[24] Y. Guo and X. J. Yang, "Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach," *Int. J. Soc. Robot.*, pp. 1–11, 2020.

[25] M. Fooladi Mahani, L. Jiang, and Y. Wang, "A Bayesian Trust Inference Model for Human-Multi-Robot Teams," *Int. J. of Soc. Robot.*, pp. 1–15, 2020.

[26] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *8th Int. Conf. Human-Robot Interact.* IEEE, 2013, pp. 251–258.

[27] J. Lee, J. Fong, B. C. Kok, and H. Soh, "Getting to Know One Another: Calibrating Intent, Capabilities and Trust for Human-Robot Collaboration," in *2020 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. IEEE, 2020, pp. 6296–6303.

[28] Y. Wang, Z. Shi, C. Wang, and F. Zhang, "Human-robot mutual trust in (semi) autonomous underwater robots," in *Cooperative Robots and Sensor Networks 2014*. Berlin, Germany: Springer, 2014, pp. 115–137.

[29] M. Patacchiola and A. Cangelosi, "A developmental Bayesian model of trust in artificial cognitive systems," in *2016 IEEE Int. Conf. Dev. Learn. Epigen. Robot. (ICDL-EpiRob)*. IEEE, 2016, pp. 117–123.

[30] S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi, "Would a robot trust you? Developmental robotics model of trust and theory of mind," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 374, no. 1771, 2019.

[31] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. 2014 Conf. Empir. Meth. Nat. Lang. Proc. (EMNLP)*, 2014, pp. 1532–1543.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8024–8035.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. Learn. Representat.*, 2015.