OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations

Anqi Xu
McGill University
Montreal, Canada
anqixu@cim.mcgill.ca

Gregory Dudek
McGill University
Montreal, Canada
dudek@cim.mcgill.ca

ABSTRACT

We present OPTIMo: an Online Probabilistic Trust Inference Model for quantifying the degree of trust that a human supervisor has in an autonomous robot "worker". Represented as a Dynamic Bayesian Network, OPTIMo infers beliefs over the human's moment-to-moment latent trust states, based on the history of observed interaction experiences. A separate model instance is trained on each user's experiences, leading to an interpretable and personalized characterization of that operator's behaviors and attitudes. Using datasets collected from an interaction study with a large group of roboticists, we empirically assess OPTIMo's performance under a broad range of configurations. These evaluation results highlight OPTIMo's advances in both prediction accuracy and responsiveness over several existing trust models. This accurate and near real-time human-robot trust measure makes possible the development of autonomous robots that can adapt their behaviors dynamically, to actively seek greater trust and greater efficiency within future human-robot collaborations.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*software psychology*

General Terms

Algorithms, Experimentation

Keywords

Trust; Dynamic Bayesian Network

1. INTRODUCTION

Trust – one's belief in the competence and reliability of another – is the cornerstone of all long-lasting collaborations. We present the design, development, and evaluation of a personalized dynamics model for a human operator's degree of trust in a robot teammate. This work focuses on asymmetric human-robot teams, where the operator occasionally intervenes to aid the autonomous robot at its given task. We conducted a large observational study with diverse roboticists belonging to a nation-wide field robotics network, to gather both empirical data as well as pragmatic insights towards the development of our trust model. The resulting Online Probabilistic Trust Inference Model (OPTIMo) formulates Bayesian beliefs over the human's moment-to-moment trust states, based on the robot's task performance and the operator's reactions over time. Our empirical analyses demonstrate OPTIMo's great performance in a broad range of settings, and also highlight improvements in both accuracy and responsiveness over existing trust models.

The end-goal of our research is to develop *trust-seeking adaptive robots*: these robots will be able to sense when the human has low trust, and adapt their behaviors in response to improve task performance and seek greater trust. In this work, we tackle an essential component of this trust-seeking methodology, by developing a trust model that can accurately and responsively quantify the human's trust states during interactions. Our ongoing research aims to integrate this online trust measure with our interactive robot behavior adaptation methods [15], towards the ultimate vision of synergistic and trust-maximizing human-robot teams.

Our contribution, OPTIMo, is a personalized model that can accurately infer human-robot trust states at various interaction time scales. OPTIMo has the unique ability to estimate the operator's trust in near real-time, whereas existing models operate on the orders of minutes or longer [7, 2, 14]. Also, OPTIMo combines the two dominant modeling approaches in the literature: applying *causal* reasoning to update the robot's deserved trustworthiness based on its task performance, and using *evidence* from direct experiences to describe a human's actual amount of trust. Our empirical results substantiate OPTIMo's diverse utilities, through accurate and responsive predictions of each user's trust states and trust-dependent behaviors, and with interpretable characterizations of the operator's trust tendencies.

2. BACKGROUND

Trust is a rich and multi-faceted construct, studied across many disciplines [8, 5], given its critical role in healthy and effective human relationships. This section explores dominant aspects of trust applicable to human-robot collaborations, and in particular to asymmetric teams. We highlight important assumptions about human-robot trust, towards establishing a coherent mathematical model.

HRI'15, March 02 - 05, 2015, Portland, OR, USA.

http://dx.doi.org/10.1145/2696454.2696492.

2.1 Trust Characterization

Lee and See surveyed the many dimensions used to characterize the basis of trust in automation [8]. These broadly fall under two categories: factors based on the automation's task *performance*, and those based on its honest *integrity*. As is typical in robotics research [2, 14, 10, 11], our work assumes that integrity-centric bases are given, i.e. our robots are obedient and never deceptive. *We henceforth assume a performance-centric definition of trust, namely one that relates solely to the robot's task efficiency and competence.*

Many representations have been proposed to quantify the degree of trust in a robot or a piece of automation. These include binary [3] and continuous [7] measures that characterize the robot's trustworthiness *caused* by its task performance, as well as ordinal scales [9, 4] used to elicit *evidence* of a person's actual amount of trust. This work incorporates both the causal and evidential modeling approaches in the literature. We employ a continuous, interval trust representation, spanning between complete distrust and absolute trust. Our model further quantifies the uncertainty in its estimated trust states, by maintaining Bayesian beliefs over the human's moment-to-moment degrees of trust.

Multiple studies *described* human-robot trust through correlations to interaction experiences and subjective assessments [9, 4], although few are capable of *predicting* a human's trust state. Lee and Moray presented a temporal model for relating trust assessments to task performance factors in a human-automation context, using an Auto-Regressive and Moving Average Value regression approach (ARMAV) [7]. We proposed a similar linear model in recent work [14] to predict event-based *changes in trust*, by relating to experience factors such as the robot's task failure rate and the frequency of human interventions to correct these failures. Desai and Yanco [2] conducted a series of robotic search and rescue experiments, during which users were asked to regularly report whether their trust has increased, decreased, or remained unchanged. These signals were quantified as $\{+1, -1, 0\}$ and integrated over time to obtain the Area Under Trust Curve (AUTC) measure. We will compare OPTIMo against each of these existing trust models in Sec. 5.3.

2.2 Interaction Context

We focus on asymmetric, supervisor-worker style human-robot teams, in which an autonomous robot "worker" is chiefly responsible for handling an assigned task. The human "supervisor" has the ability to intervene and take over control, but should do so only when necessary, for example to correct the robot's mistakes, or to switch to a new task objective. The human's interventions are assumed to always take precedence over the robot's autonomous controller.

The trust model developed in this work generalizes readily to human-robot teams in arbitrary domains. Nevertheless, our work focuses on visual navigation contexts, where an aerial robot is autonomously driven by an adaptive visual boundary tracking algorithm [15]. This robot can learn to follow diverse terrain contours, such as the coastline target shown in Fig. 1. We investigate scenarios where an operator collaborates with this robot, and model the evolution of the human's trust in the robot's abilities to reliably follow the designated boundary targets. Visual navigation tasks are appealing because humans innately excel at them, whereas the necessary complexity in autonomous solutions [2, 15] leads to uncertainty, and thus warrants the need for trust.

Figure 1: Live camera feed from an aerial robot overlaid with an autonomous boundary tracker's steering command (blue arrow), and the human's interventions (green arrow). Additional overlays are pertinent to our observational study (see Sec. 3).

3. METHODOLOGY

Our performance-centric trust modeling approach is predicated on two simple observations of asymmetric human-robot collaborations. Firstly, the robot's trustworthiness arises due to its task performance: efficient progress lead to greater trust, whereas low reliability induces losses in trust. Secondly, whenever the human intervenes, it often reflects a lapse in trust due to the robot's task failures.

Formally, the goal of this work is to estimate the degree of trust $t_k \in [0, 1]$ felt by a human towards a robot, at time steps $k = 1{:}\, K$ during their interactions. We tackle this problem by relating the human's latent trust states to observable factors in the interaction experience. In particular, the well-studied link between trust and the robot's task performance p is often quantified through the failure rate of its autonomous system [7, 2]. In addition, human interventions i (i.e. toggles between autonomous and manual control modes) are known to be strong evidential predictors for trust [2, 14]. We further consider extrinsic factors e that cause the operator to intervene irrespective of trust, for instance when steering along a new terrain contour to train our boundary-tracking robot to follow this updated target.

We conducted an observational study to collect interaction experiences towards modeling trust relationships. This study further yielded pragmatic insights about trust and its related constructs, for asymmetric human-robot teams.

3.1 Robot Interface

While supervising the boundary-tracking aerial robot, the human operator is presented with an interface showing the onboard camera feed, as seen in Fig. 1. The operator can take over control of the vehicle at any time, by moving an analog stick on a gamepad. Even during these periods of manual interventions, the boundary tracker continues to process the camera stream and displays its generated headings. This feedback aids the human in deciding if the robot is capable of performing the task on its own, or if further assistance is needed when the tracker is behaving poorly.

3.2 Trust Assessment Elicitation

In addition to logging performance and intervention factors that are available in typical interaction experiences, our study also queried the human's trust assessments occasionally. These factors are used to train personalized instances of our trust model, and to evaluate their prediction accuracies. Importantly, after the training period, our model can operate without such assessment data, although their availabilities will strengthen the resulting trust predictions.

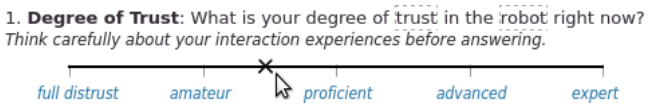


Figure 2: Post-session trust feedback questionnaire.

Our study is divided into multiple interaction sessions. After each session, the user is asked to provide an assessment of their trust state, based on the cumulative interaction experience. These trust feedback f are queried using a modified Visual Analog Scale (VAS) [12], as shown in Fig. 2. Multiple anchor points are introduced to this interval scale to reduce end-aversion bias [1]. This format also addresses concerns reported by several pilot study users regarding difficulties in quantifying their trust updates based on end-anchors only.

It is important to minimize the number of trust feedback queries, both to reduce interaction interrupts, and to mitigate added mental strain. As a separate source of trust evidence, each user is asked to report changes in their trust states during interactions, by pressing buttons on the gamepad indicating whether trust has been gained, lost, or remained unchanged, i.e. $c = \{+1, -1, 0\}$. Similar to previous studies [2], we encourage periodic reporting of c by prompting 5 seconds after each button press, both using visual feedback (with a "`t?`" icon in Fig. 1) and by vibrating the gamepad.

3.3 Observational Study Design

Our study involved two interaction scenarios that tasked the human-robot team to follow sequences of boundary targets. Each scenario is separated into 1 to 2 minute-long sessions, to enable frequent post-session queries for the user's trust feedback f. The first scenario involved patrolling along a highway stretch, a forest path, and a coastline, whereas the second scenario surveyed the hill-sides and shores of a fjord.

This study was carried out using a simulated environment to enforce repeatability, and to have ground truth in the form of the ideal target boundary paths. In order to ensure similar experiences across users, the interface provided visual warnings whenever the robot deviated away from the designated boundary. If the user failed to recover in a timely manner, the interaction then reset to a previous state.

Every time the boundary tracker processed a camera frame, we recorded whether it had failed to detect any boundaries (i.e. AI failures reflecting task performance $p \in \{0, 1\}$), and the human's intervention state $i \in \{0, 1\}$ at that time. We also noted frames in proximity to a change-over between boundary targets, which constituted extraneous intervention causes $e \in \{0, 1\}$. Furthermore, we logged button presses indicating trust changes $c \in \{-1, 0, +1\}$ as well as the user's absolute trust feedback $f \in [0, 1]$ following each session.

This study was conducted in a fully automated manner. Following a demographics survey, a short slide-show elaborated on the study and its interface. Next, the user worked through an interactive tutorial and 2 practice sessions, to familiarize with the tasks and trust feedback queries. Afterwards, a 3-session interaction scenario and another 2-session scenario were each administered *twice*, for a total of 10 recorded sessions. All users in this observational study underwent the task sessions in the same order, since the study was primarily aimed at collecting interaction experiences. This is in contrast to our previous controlled trust experiments for quantifying event-related effects [14].

3.4 Observational Study Results

The final form of the observational study resulted from iterative refinements following a series of pilot runs. We recruited 21 roboticists from 7 universities during a nationwide field robotics gathering to take part in this study. Participants were predominantly graduate students (86%), and the average age of users was 27 ($\sigma \approx 4$). A typical study run entailed 30 minutes of interaction with the boundary-tracking robot, operating at a $15\,Hz$ frame rate.

The 2 task scenarios were each administered twice to assess whether users behaved consistently in similar situations. We found no significant differences in the rate of human interventions i per matching sessions (2-tailed paired $t_{104} = 1.49, p = 0.14$). The numerical sum of reported trust changes c across paired session instances also did not reveal any significant differences (2-tailed paired $t_{104} = -0.43, p = 0.67$). In contrast, trust feedback f were found to be significantly different when users repeated the same scenarios (1-tailed paired $t_{104} = -4.85, p \ll 0.01$). We conclude that users reacted consistently to similar events, yet their trust assessments changed over time as they accumulated more interaction experiences. These results thus substantiate the need to model the *temporal dynamics* of human-robot trust.

A linear regression on trust feedback f was performed to identify significant covariates from interaction experience. An analysis of variance revealed that both the user identifier ($F_{20,188} = 17.4, p \ll 0.01$) and the ratio of user interventions i per session ($F_{1,207} = 76.2, p \ll 0.01$) were significantly related to trust feedback f, whereas the ratio of AI failures per session was related to a lesser degree ($F_{1,207} = 3.02, p = 0.08$). The strong user dependence on trust feedback supports the need for a *personalized* model of trust, which is consistent with findings in our previous study [14]. Also, the stronger trust correlation of user interventions over AI failures is captured by the structure of our dynamic Bayesian trust model, to be discussed in Sec. 4.1.

Users of our study further provided a number of qualitative remarks, reflecting vital insights about the evolution of their trust in the robot's task performance. Several users reported that their "trust changed when the robot did something unpredictable", which suggests a dependency between the trust state t_k at time k and the *change* in the robot's recent task performance, i.e. $p_k - p_{k-1}$. Others said that their "trust fluctuated a lot initially", given the lack of prior experiences with the robot. This suggests that it is sensible to assume a *uniform prior* belief on each user's initial degree of trust when interacting with a new robotic system.

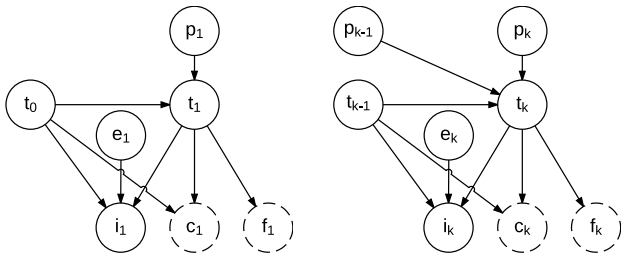
During pilot runs of the study, users frequently pressed the 'trust gained' and 'trust lost' buttons unintentionally when prompted for c. Consequently, the slide-show in the revised study explicitly encouraged participants to press the 'trust unchanged' button as a default. Nevertheless, multiple participants reported that they found it "hard to suppress the urge to hastily press 'trust gained' or 'trust lost'", and re-

called making multiple accidental misclicks. We will model this *idling bias* into the relationship between a user's latent trust state t_k and reports of trust changes c.

4. HUMAN-ROBOT TRUST MODEL

In this section, we present the Online Probabilistic Trust Inference Model (OPTIMo) for asymmetric human-robot teams. OPTIMo treats the degree of human-robot trust t_k at each time step k as a random variable, and maintains belief distributions for these performance-centric trust measures by deducing from various factors of the interaction experience. This probabilistic representation is useful for inferring the human's expected trust state at a given time, as well as the amount of uncertainty of each such estimate.

OPTIMo is represented as a Dynamic Bayesian Network [6], as shown in Fig. 3. Its graph structure efficiently encodes local relationships between trust and its related factors, as well as the evolution of trust states over time. This Bayesian model also processes variable-rate sources of information in a probabilistically sound manner, and can further accommodate an arbitrary belief for prior trust t_0. Finally, OPTIMo combines the two main approaches used by existing trust models, namely by applying *causal* reasoning to update the robot's trustworthiness based on its task performance [7, 10], and analyzing *evidential* factors to quantify each user's actual degree of trust [2, 14].



(a) first time step (b) general form

Figure 3: Dynamic Bayesian structure of the Online Probabilistic Trust Inference Model (OPTIMo). Dashed factors are not observed on all time steps k, and are not mandatory for inferring trust states t_k.

4.1 Local Trust Relationships

As seen in Fig. 3, OPTIMo relates the human's latent trust state t_k causally to the robot's task performance p. It also incorporates several sources of trust-induced evidence, such as user interventions i, trust change reports c, and absolute trust feedback f. This model processes a continuous period of interaction as a sequence of K non-overlapping time windows, $k = 1{:}\,K$, each lasting W seconds. We define the window-aggregated state of task performance, $p_k \in [0, 1]$, as the ratio of frames within the k-th time window for which the robot's autonomous controller failed to produce any commands. Similarly, $i_k \in \{0, 1\}$ reflects whether the operator had intervened or not during its time window k. The extraneous cause state $e_k \in \{0, 1\}$ records the presence of a change in task targets, and is connected as a parent link to i_k. Finally, sign-aggregated trust change reports $c_k \in \{-1, 0, +1, \varnothing\}$ and each individual trust feedback $f_k \in \{[0, 1], \varnothing\}$ help to ground the model's estimates for latent trust states t_k as further evidence. Since these trust assessments c_k, f_k occur sporadically, $\varnothing$ is used to denote their non-occurrences at various times.

Links to each factor in this discriminative Bayesian model are quantified as a conditional probability distribution (CPD). The amount for which trust t_k is expected to change given the robot's recent and current task performances, p_{k-1}, p_k, is reflected by a linear Gaussian CPD:

$$\begin{aligned}\mathcal{P}(t_k,t_{k-1},p_k,p_{k-1}) &:= Prob(t_k|t_{k-1},p_k,p_{k-1}) \approx {}^{*} \\ &\mathcal{N}\left(t_k;\ t_{k-1} + \omega_{tb} + \omega_{tp}\, p_k + \omega_{td}\ (p_k - p_{k-1})\,,\ \sigma_t\right)\end{aligned} \tag{1}$$

where $\mathcal{N}(x;\ \mu, \sigma) := \frac{1}{\sqrt{2\pi}\sigma} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ denotes a Gaussian distribution for the random variable x, with mean μ and standard deviation σ. The expression for this Gaussian CPD's mean represents the expected update to trust t_k from its previous state t_{k-1}, as a weighted sum of a bias term (i.e. propensity to trust), as well as performance-induced factors. The personalized parameters $\omega_{t\mathbf{b}}, \omega_{t\mathbf{p}}, \omega_{t\mathbf{d}}$ reflect the relative impacts on each user's trust updates of the **b**ias, the current task **p**erformance, and the **d**ifference in the robot's performance. The propagation uncertainty parameter σ_t quantifies the variability in each user's trust update dynamics.

OPTIMo explains the probability of interventions i_k by its diverse causes, such as the current trust state t_k, a recent change in trust $\Delta t_k := t_k - t_{k-1}$, and extraneous factors e_k. These linkages are modeled as a logistic CPD:

$$\begin{aligned}\mathcal{O}_i(t_k = 1,t_{k-1}, i_k, e_k) &:= Prob(i_k = 1|t_k, t_{k-1}, e_k) = \\ &\mathcal{S}\left(\omega_{ib} + \omega_{it}\, t_k + \omega_{id}\, \Delta t_k + \omega_{ie}\, e_k\right) \\ Prob(i_k = 0|t_k, t_{k-1}, e_k) &= 1 - Prob(i_k = 1|t_k, t_{k-1}, e_k)\end{aligned} \tag{2}$$

where $\mathcal{S}(x) := (1 + exp(-x))^{-1}$ is the sigmoid distribution for the binary random variable x. The parameters $\omega_{ib}, \omega_{it}, \omega_{id}, \omega_{ie}$ quantify the bias and weights of the various causes explaining the intervention state i_k.

During time steps when the user reports trust changes $c_k \in \{-1, 0, +1\}$, these are accounted as evidence to ground the latest update to latent trust, Δt_k. Reports of 'trust gains' and 'trust losses' are modeled as sigmoid CPDs:

$$\begin{aligned}\mathcal{O}_c(t_k, t_{k-1}, c_k) :=& Prob(c_k|t_k, t_{k-1}) \\ Prob(c_k = +1|t_k, t_{k-1}) =& \beta_c + (1 - 3\beta_c) \cdot \mathcal{S}\left(\kappa_c\left[\Delta t_k - o_c\right]\right) \\ Prob(c_k = -1|t_k, t_{k-1}) =& \beta_c + (1 - 3\beta_c) \cdot \mathcal{S}\left(\kappa_c\left[-\Delta t_k - o_c\right]\right) \\ Prob(c_k = 0|t_k, t_{k-1}) =& 1 - Prob(c_k = +1|t_k, t_{k-1}) \\ & - Prob(c_k = -1|t_k, t_{k-1})\end{aligned} \tag{3}$$

This CPD parameterizes the nominal offset o_c in a change to latent trust Δt_k that is required to cause the user to report a non-zero c_k, along with the variability κ_c in the reporting likelihoods. In addition, the uniform error term β_c captures the *idling bias* observed during our study, where users sometimes reported erroneous trust changes when prompted.

Finally, OPTIMo uses a zero-mean Gaussian CPD to quantify the uncertainty σ_f in each user's absolute trust feedback f_k with respect to their true latent trust state t_k:

$$\mathcal{O}_f(t_k, f_k) := Prob(f_k|t_k) \approx {}^{\dagger}\mathcal{N}\left(f_k;\ t_k, \sigma_f\right) \tag{4}$$

*Since $t_k \in [0, 1]$ is bounded, the cumulative Gaussian densities below $t_k = 0$ and above $t_k = 1$ must be added to these end-states, enforcing a proper probability density function.

†Akin to $\mathcal{P}(t_k, t_{k-1}, p_k, p_{k-1})$, cumulative densities beyond the range of $f_k \in [0, 1]$ must be added to boundary states.

All of these trust relationships have been corroborated by prior literature [7, 2, 14], and were further supported by results of our observational study. In particular, analyses in Sec. 3.4 found that user interventions i were much more strongly correlated to trust than the robot's task performance p. Consequently, this temporal Bayesian model uses p_k to causally propagate the trust belief t_k to a set of *plausible* next states, and then uses trust-induced evidences i_k, c_k, and f_k to *exclude* inconsistent hypotheses. This graph structure also arises naturally as the causal depiction of the human's trust-driven decision process.

4.2 Inference, Personalization, and Prediction

This model can be used to estimate the probabilistic belief that the user's trust state $t_k \in [0, 1]$ at time step k takes on any particular value. Trust inference can be carried out in 2 different contexts: firstly, given a history of past experience data, we can query the *filtered belief* at the *current* time k, $bel_f(t_k) = Prob(t_k|p_{1:k}, i_{1:k}, e_{1:k}, c_{1:k}, f_{1:k}, t_0)$. We can also compute the *smoothed belief* at *any* time step $k \in [0\colon K]$ within a previously recorded interaction dataset with K time steps, $bel_s(t_k) = Prob(t_k|p_{1:K}, i_{1:K}, e_{1:K}, c_{1:K}, f_{1:K}, t_0)$. Both types of belief inferences can be derived from OPTIMo's graph structure [13]:

$$\overline{bel}(t_k, t_{k-1}) := \mathcal{O}(t_k, t_{k-1}, i_k, e_k, c_k, f_k) \cdot \mathcal{P}(t_k, t_{k-1}, p_k, p_{k-1}) \cdot bel_f(t_{k-1}) \quad (5)$$

$$bel_f(t_k) = \frac{\int \overline{bel}(t_k, t_{k-1})\ dt_{k-1}}{\iint \overline{bel}(t_k, t_{k-1})\ dt_{k-1} dt_k} \quad (6)$$

$$bel_s(t_{k-1}) = \int \frac{\overline{bel}(t_k, t_{k-1})}{\int \overline{bel}(t_k, t_{k-1})\ dt_{k-1}} \cdot bel_s(t_k)\ dt_k \quad (7)$$

where $\mathcal{O}(t_k, t_{k-1}, i_k, e_k, c_k, f_k)$ denotes the product of one or more observation CPDs $\mathcal{O}_i(\cdot), \mathcal{O}_c(\cdot), \mathcal{O}_f(\cdot)$ at each time step k, based on whether $c_k \neq \varnothing$ and $f_k \neq \varnothing$. We also assume a uniform prior trust belief, $bel_f(t_0) := Prob(t_0) = 1$, when users begin to interact with a new robot.

It is worthwhile to note the efficient linear-time complexities of the filtering and smoothing algorithms. The filtered belief $bel_f(t_k)$ at time k in Eqn. 6 is updated recursively from its previous belief distribution, $bel_f(t_{k-1})$, in Eqn. 5; a similar recursive relationship also holds for the smoothed belief $bel_s(t_k)$ in Eqn. 7. Thus, starting from a given prior trust belief $bel_f(t_0)$, one can compute filtered beliefs $bel_f(t_k)$ forward in time in a single pass, and then also ascertain smoothed beliefs $bel_s(t_k)$ backwards in time sequentially.

In order to personalize OPTIMo to a particular user's behaviors and trust tendencies, we use the hard-assignment Expectation-Maximization (EM) algorithm [6] to find optimized model parameters $\mathbf{\Theta}^*$ (e.g. $\omega_{tb}, \sigma_t, ...$) given a *training set* of interaction experiences. Hard-assignment EM jointly optimizes the observational likelihood of all interaction data and the most likely sequence of latent trust states, as follows:

$$\mathbf{\Theta}^* = \arg\max_{\mathbf{\Theta}} \max_{t_{1:K}} Prob(t_{1:K}, p_{1:K}, i_{1:K}, e_{1:K}, c_{1:K}, f_{1:K}|t_0)$$

In addition to inferring trust beliefs, OPTIMo can also be used to predict probability distributions for the various evidential factors i_k, c_k, f_k. Using the intervention factor as an example, we can remove all instances $i_{1:k}$, and then predict the likelihood of observing a particular i_k state as:

$$Prob(i_k|p_{1:k}, e_{1:k}, c_{1:k}, f_{1:k}) = \frac{\iint \mathcal{O}_i(t_k, t_{k-1}, i_k, e_k) \cdot \overline{bel}(t_k, t_{k-1})\ dt_k\ dt_{k-1}}{\iint \overline{bel}(t_k, t_{k-1})\ dt_k\ dt_{k-1}} \quad (8)$$

Derivations and examples of filtering, smoothing, and prediction for Dynamic Bayesian Networks are presented in [13].

4.3 Histogram Inference Engine

We implemented OPTIMo using a histogram-based inference method, which approximates continuous belief densities as discrete vectors of probability masses. Specifically, we discretized the trust range $[0, 1]$ into B equally-spaced bins, and approximated the likelihood of t_k taking on a particular value τ as the probability mass of the nearest bin center. The precision of this histogram approximation improves when using larger bin sizes B, at the cost of additional computations. A similar approximation is used when predicting trust feedback, where the distribution for $f_k \in [0, 1]$ is discretized into 100 bins, and each bin center's probability mass is computed using a form similar to Eqn. 8.

We personalized model instances using EM with multiple restarts, to avoid convergence to local optima. In each EM run, model parameters $\mathbf{\Theta}$ are initiated from hand-selected or random values, and then iteratively improved using constrained least squares optimization. An EM run is terminated when parameters have stabilized within expected tolerances, or after a maximum number of iterations have lapsed. Further algorithmic details and examples of histogram inference and model training for Dynamic Bayesian Networks are discussed in [13].

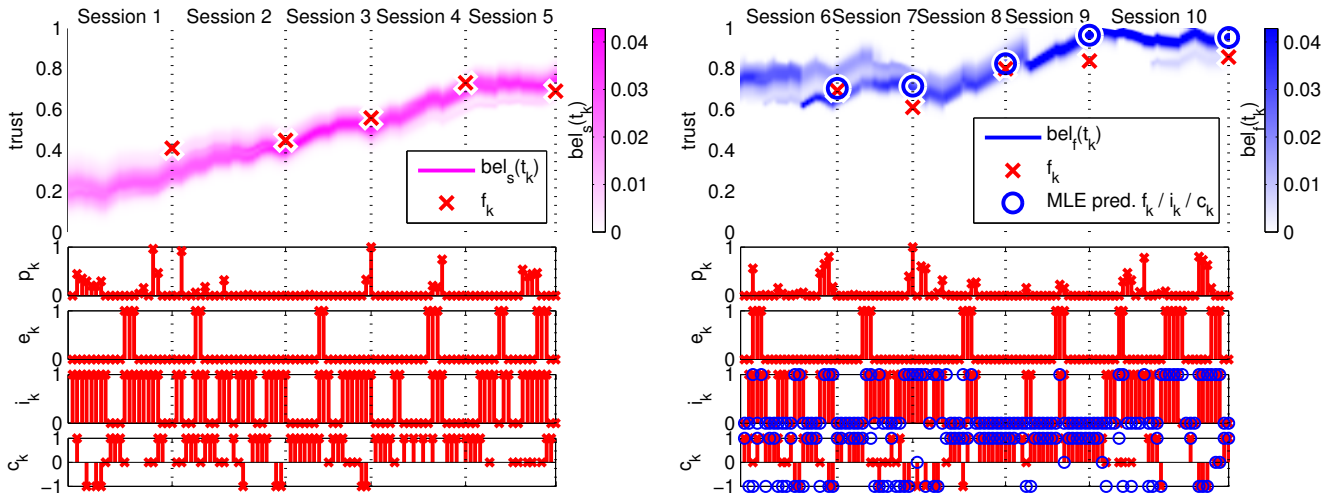
5. EVALUATION OF TRUST MODEL

This section describes empirical assessments of OPTIMo under diverse settings for its non-trainable parameters, and compares it to several existing temporal trust models. Since our approach assumes that the human's trust state t_k is latent and thus never observable, we quantify the performance of each model by its ability to predict trust-induced behaviors and assessments. These include user interventions i_k, trust change reports c_k, and absolute trust feedback f_k.

Each evaluation run begins by aggregating raw interaction datasets into W-second time windows. We personalize OPTIMo instances using *training sets* consisting of each user's experiences during the first 5 study sessions, while assuming uniform prior trust. After optimizing model parameters to capture each operator's trust tendencies, we compute the filtered trust belief at the end of the 5 training sessions.

This trust belief is used as the prior distribution for the *test set*, comprising of the remaining 5 repeated sessions. We conduct separate prediction assessments for $i_{1:k}$, $c_{1:k}$, and $f_{1:k}$: for each target variable, first all of its instances in the test set are removed, and predicted likelihoods are computed for each omitted occurrence, following the form of Eqn. 8. We then compute Maximum Likelihood Estimates (MLE) using the predicted beliefs, and compare the resulting values against the omitted observed data, $i_{1:k}$, $c_{1:k}$, or $f_{1:k}$.

The outcomes of each model evaluation process consist of prediction accuracies for interventions, $acc_i \in [0\%, 100\%]$ accuracies for trust change predictions acc_c, and the root-mean-squared-error for absolute trust feedback, $RMSE_f$. We also compute the Pearson product-moment correlation



(a) training set sessions

(b) test set sessions

Figure 4: Inferences and predictions for a typical trained OPTIMo instance ($W = 3\,s$, $B = 300$ bins, $\sigma_f = 0.1$).

coefficient ρ between matching sets of observed and predicted trust feedback f_k. This metric is used to compare OPTIMo against other models that quantify trust on a different scale. All evaluation metrics presented in this section (except Sec. 5.1) are aggregated across all 21 users' results.

A few of OPTIMo's settings cannot be optimized using interaction experiences, yet they affect the model's performance. The window duration W determines the time scale of interaction, and also reflects the trust inference latency for online model queries. The number of histogram bins B affects the precision of the discrete approximation to the underlying continuous trust beliefs. Finally, the uncertainty σ_f in each user's trust feedback f_k captures the relative influence of f_k on the inferred beliefs for latent trust t_k. Sec. 5.2 will investigate the effects of each of these non-trainable parameters on the resulting models' performance.

5.1 Characteristics of a Trained Model

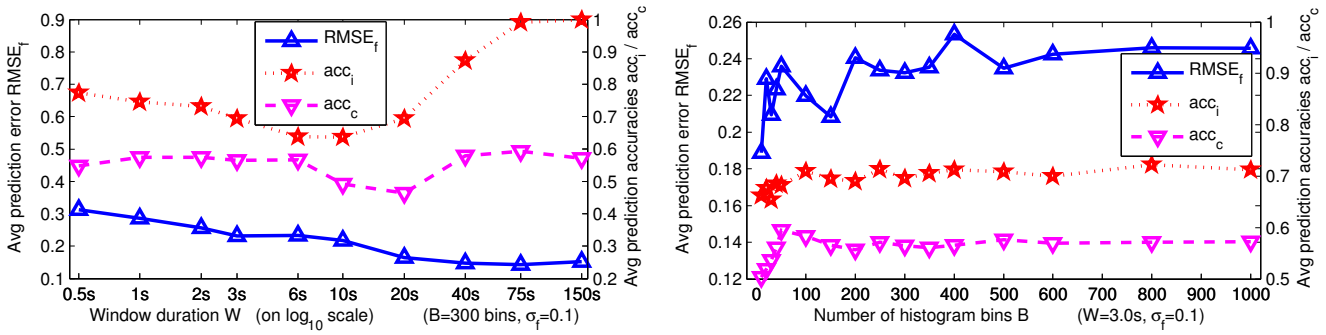
We begin by highlighting several features of a typical personalized OPTIMo instance under sensible settings: $W = 3\,s$, $B = 300$ bins, $\sigma_f = 0.1$. Fig. 4(a) depicts the Bayesian-*smoothed* trust beliefs $bel_s(t_k)$ during the training sessions, after parameters $\boldsymbol{\Theta}$ of the model have been optimized to best match the training dataset. Fig. 4(b) shows the Bayesian-*filtered* beliefs $bel_f(t)$ inferred during the test sessions, as well as MLE predictions (shown as blue circles) for trust feedback f_k, user interventions i_k, and trust change reports c_k. The switch to *filtered* trust beliefs ensures that model evaluation is carried out in an online manner, as if predictions were obtained live during the latter 5 sessions.

The inferred trust beliefs, depicted as vertical color-coded slices in Fig. 4, reflect a precise characterization of this specific user's trust tendencies. This can be seen from the accurate prediction results for various evidence factors in the test set: $RMSE_f = 0.09, acc_i = 72.41\%, acc_c = 70.00\%$. Also, despite the small number of trust feedback (e.g. 5 in each of the training and test sets), test-set predictions of f_k yielded highly significant correlations $\rho = 0.91$ ($p < 0.01$). Furthermore, the visualization of the inferred trust beliefs highlights OPTIMo's unique ability in characterizing *multi-modal distributions* for the user's temporal trust states. These competing trust hypotheses arise when the human's actions contradict recent trust reports, or when the user's reactions are notably different from prior interaction experiences.

This OPTIMo instance is personalized through trust propagation settings $\omega_{tb} = 0.0064, \omega_{tp} = -0.0153, \omega_{td} = -0.0029$, cause-of-intervention weights $\omega_{ib} = 131.2, \omega_{it} = -157.1$, $\omega_{id} = -9887, \omega_{ie} = 83.84$, and trust changes reporting traits $o_c = 0.0003, \kappa_c = 1277, \beta_c = 1.063 \times 10^{-7}$. These settings optimize the joint likelihood of the diverse interaction factors in the training set, and can also be interpreted to quantify the user's trust tendencies. Specifically, the large cause-of-intervention weights suggest that this operator does not intervene at a maximal trusting state, but will most certainly take over control for medium-to-low trust states ($t_k < 0.8$), when trust drops by even 0.004, or to address extraneous factors such as intentional changes to the task target ($e_k = 1$). Also, the trust propagation settings indicate that between consecutive time steps (of $W = 3\,s$), the user's trust state increases nominally by 0.0064 if the task performance is good, drops by 0.0118 initially upon an AI failure, and continues to decrease by 0.0089 throughout a contiguous period of failures. We deduce that this operator penalizes robot failures with 38% more trust loss, when compared to trust gains during competent operations. This quantifies the common adage that "it is easy to lose trust, but hard to regain it".

5.2 Effects of Non-Trainable Parameters

Fig. 5(a) illustrates the effects of the time window duration W on OPTIMo's performance. The prediction errors for trust feedback, $RMSE_f$, consistently decrease as the window duration widens. This trend arises due to having more frequent f_k observations at coarser time scales, which allows the trained model to more accurately predict f_k through the latent trust state t_k. In contrast, prediction accuracies for i_k and c_k drop slightly as W increases from $0.5\,s$ to $20\,s$, since occurrences of per-frame i and per-report c values are collapsed into fewer window-aggregated factors at coarser time scales. The high prediction accuracies at session-level time scales of $W = 150\,s$ is similarly explained by statistical degeneracy caused by very few samples of i_k and c_k, although even in these cases the model yields excellent performance.



(a) time window duration W

(b) number of histogram bins B

Figure 5: Effects of non-trainable model parameters on the prediction performance of OPTIMo.

The opposite effects that W has on predictions for f_k, versus those for i_k and c_k, supports the contrast between absolute trust assessments for summarizing *cumulative* experiences, and the *reactive* natures of user interventions and trust change reports. The fact that our results capture this contrast demonstrates OPTIMo's versatility. Furthermore, despite having poor $RMSE_f$ for sub-second time windows (e.g. $W = 0.5\,s$), competent prediction accuracies for i_k and c_k reflect the usefulness of OPTIMo's inferred trust beliefs at these extremely fine time scales. This degree of prediction responsiveness is unseen in existing trust models, which operate at much coarser scales of minutes or longer [7, 2, 14].

Fig. 5(b) shows the effects of using histograms with different bin sizes B. Prediction results are varied among small bin sizes ($B < 200$). These unreliable accuracies are likely caused by under-sampling of the underlying continuous trust beliefs. For example, the trained model in Sec. 5.1 indicates that the user's latent trust t_k increases nominally by 0.0064 between time steps under typical operations; this requires at least $B > 156$ bins to represent faithfully. Beyond these under-sampling errors, our empirical results indicate that using many bins does not lead to greater performance. We therefore conclude that a histogram with $B = 300$ bins sensibly captures beliefs of a typical robotict's trust dynamics.

In general, it is difficult to estimate the amount of variability of a human's reports on an unmeasurable sentiment such as trust. Since we employed a 5-anchors format to elicit trust feedback $t_k \in [0, 1]$ (see Fig. 2), a conservative estimate of variability is on the scale of consecutive anchor points, e.g. $\sigma_f \approx 0.1$. We trained model instances with varying variability values σ_f, and found that this parameter had minimal effects on OPTIMo's performance. For instance, under reasonable settings $W = 3\,s, B = 300$ bins, results assuming extremely precise trust feedback ($\sigma_f = 0.001$), $acc_i = 69.43\%$, $acc_c = 51.52\%, RMSE_f = 0.21$, are not noticeably different from performance assuming severely unreliable assessments ($\sigma_f = 0.3$), $acc_i = 71.41\%, acc_c = 55.81\%, RMSE_f = 0.29$.

5.3 Comparison with Existing Trust Models

We contrasted OPTIMo's trust prediction performance against those of several existing models. These include the Auto-Regressive Moving Average Value (ARMAV) model for quantifying human-automation trust [7], a stepwise regression model for predicting *changes* in trust (dTrust) [14], and the Area Under Trust Curve (AUTC) metric [2]. We compared these models against two OPTIMo variants ($B =$ 300 bins, $\sigma_f = 0.1$) at different time scales, i.e. $OPTIMo_{fine}$ with $W = 3\,s$, and $OPTIMo_{coarse}$ with $W = 150\,s$.

The ARMAV model [7] associates the degree of trustworthiness in an automated system as a linear function of its task performance and internal failure rate, at the current and last time steps. We implemented three variants of this first-order lag system, including $ARMAV_{real}$, which solely considers recent AI failure rates p_k, p_{k-1}, and $ARMAV_{perf}$, which further adds an external task performance metric, measured as the ratio of frames for which the target boundary was completely out of the robot's view. This latter variant is expected to be more accurate, although it requires a performance metric that is typically not available for many families of tasks (and is thus not used by OPTIMo). Both model variants were personalized using interaction data from the 5 training sessions, similar to OPTIMo's training process. We further computed a user-aggregated regression variant, $ARMAV_{aggr}$, using both AI failure and external performance data from *all users'* training sessions, to faithfully replicate the form used by the original authors.

In prior work [14] we derived a similar linear regression model, but for predicting changes in trust caused by different interaction events. In addition to linking trust to AI failures p_k, this dTrust model also relates trust to user interventions i_k, and to the robot's instantaneous task performance, which is represented as the distance from the ground truth target trajectory. Since this metric is typically not available during online operations, we computed per-user regression variants that both ignored ($dTrust_{real}$) and included ($dTrust_{gt}$) this ground truth data. Furthermore, a user-aggregated variant $dTrust_{aggr}$ incorporates interaction experiences from all users, and best reflects the originally-proposed model form. To evaluate these models, we computed all predicted trust changes for the 5 test sessions, and integrated the updated trust states starting from the final trust feedback at the end of the training sessions.

In the experiments conducted by Desai and Yanco [2], users were asked to report their trust changes $c \in \{-1, 0, +1\}$ at regular intervals through button presses. The cumulative sums of these values were used to characterize trust states at different times, and is termed the Area Under Trust Curve (AUTC) metric. We computed AUTC values at the end of each test session, and compared Pearson's ρ correlations between these unscaled predictions and trust feedback f_k.

Table 1 summarizes RMSE and Pearson's ρ statistics for trust predictions for the various models. OPTIMo yielded prediction accuracies and correlations comparable to the

Table 1: Model comparison of trust prediction errors and Pearson's ρ. Dark & light shades highlight best and second best results for each metric. Statistics for personalized models were across all users.

	avg (std) $RMSE$	% $sign\ \rho$ ($\alpha < 0.10$)	avg (std) $sign\ RMSE$	avg $sign\ \rho$
$OPTIMo_{fine}$	0.13 (0.12)	33.33%	0.11 (0.14)	0.88
$OPTIMo_{coarse}$	0.10 (0.09)	33.33%	0.09 (0.10)	0.90
$ARMAV_{real}$ [7]	0.32 (0.42)	9.52%	0.52 (0.05)	0.86
$ARMAV_{perf}$ [7]	0.27 (0.21)	9.52%	0.12 (0.04)	0.85
$ARMAV_{aggr}$ [7]	0.14	—	—	0.64
$dTrust_{real}$ [14]	0.76 (1.68)	9.52%	0.17 (0.10)	0.92
$dTrust_{gt}$ [14]	0.36 (0.32)	14.29%	0.17 (0.07)	0.93
$dTrust_{aggr}$ [14]	0.19	—	—	0.73
$AUTC$ [2]	—	38.10%	—	0.90

best variants of all existing trust models. Focusing on a session-wide scale, $OPTIMo_{coarse}$ produced notably more accurate trust predictions against other models using similar information, i.e. $ARMAV_{real}$, $dTrust_{real}$, $AUTC$. Even when other models incorporated additional sources of data, OPTIMo's performance remained highly competitive.

These results show that OPTIMo is able to infer the user's trust states with greater fidelity than existing methods, and at much finer time scales, on the order of seconds. This is achieved by relating the user's *latent* trust state to *observed* factors of the interaction experience. OPTIMo's use of a probabilistic trust representation also captures additional state information, such as multi-modal hypotheses and the uncertainty in trust estimates. The fact that trained OPTIMo instances accurately predict trust-induced behaviors and attitudes in near real-time, without requiring users to tediously provide trust feedback every few seconds, highlights the unique value of this online human-robot trust model.

6. CONCLUSION

We introduced the Online Probabilistic Trust Inference Model (OPTIMo): a personalized *performance-centric* trust model that is capable of inferring a human operator's degree of trust in an autonomous robot. OPTIMo incorporates the two dominant modeling approaches in the literature, namely through causal reasoning of the robot's trustworthiness given its task performance, and using evidence from interaction data to support beliefs about the human's latent trust state. We conducted an observational study on a large group of roboticists, and collected a substantive dataset as well as valuable insights that helped in shaping OPTIMo's structure. We demonstrated success at accurately predicting trust-induced behaviors of human operators while collaborating with an autonomous boundary following robot, although OPTIMo's generic form can be scaled and instantiated to suit other task contexts as well. Our empirical analyses extensively quantified the strong performance of many variants of this human-robot trust model. Our results further showed that OPTIMo can predict trust assessments with greater accuracies and at much finer time scales compared to existing works. These findings highlight the importance and uniqueness of OPTIMo towards developing responsive *trust-seeking adaptive robots*.

In ongoing work we are studying the use of OPTIMo for grouping like-minded operators. This would allow pretrained models to predict trust states of novel users, while requiring minimal amounts of personalization data. We are also investigating the use of Monte Carlo approximate inference techniques, with the aim of further speeding up the trust inference process. Finally, we have begun integrating OPTIMo with our interactive robot behavior adaptation methods [15], towards our research end-goal of building robots that can actively seek to maximize the user's trust.

7. ACKNOWLEDGMENTS

We would like to thank all participants of our observational study, which was sanctioned by McGill University's Research Ethics Board (#183-1112). This work was funded by the NSERC Canadian Field Robotics Network (NCFRN).

8. REFERENCES

[1] J. A. Cowley and H. Youngblood. Subjective response differences between visual analogue, ordinal and hybrid response scales. *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, 53(25), 2009.

[2] M. Desai. *Modeling Trust to Improve Human-Robot Interaction*. PhD thesis, Computer Science Department, University of Massachusetts Lowell, 2012.

[3] R. J. Hall. Trusting your assistant. In *Knowledge-Based Soft. Eng. Conf. (KBSE'11)*, 1996.

[4] J.-Y. Jian, A. M. Bisantz, and C. G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International J. of Cognitive Ergonomics*, 4(1), 2000.

[5] A. Jøsang, R. Hayward, and S. Pope. Trust network analysis with subjective logic. In *Australasian Computer Science Conf. (ACSC'06)*, 2006.

[6] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

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

[8] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 2004.

[9] B. M. Muir. *Operators' trust in and use of automatic controllers in a supervisory process control task*. PhD thesis, University of Toronto, 1989.

[10] A. Pierson and M. Schwager. Adaptive inter-robot trust for robust multi-robot sensor coverage. In *Int. Sym. on Robotics Research (ISRR'13)*, 2013.

[11] C. Pippin and H. I. Christensen. Trust modeling in multi-robot patrolling. In *Proc. of the IEEE Int. Conf. on Rob. and Auto. (ICRA'14)*, 2014.

[12] U.-D. Reips and F. Funke. Interval-level measurement with visual analogue scales in internet-based research: Vas generator. *Behavior Research Methods*, 2008.

[13] A. Xu. 2-Step Temporal Bayesian Networks (2TBN): filtering, smoothing, and beyond. Technical Report TRCIM1030, McGill U., 2014. www.cim.mcgill.ca/~anqixu/pub/2TBN.TRCIM1030.pdf.

[14] A. Xu and G. Dudek. Towards modeling real-time trust in asymmetric human-robot collaborations. In *Int. Sym. on Robotics Research (ISRR'13)*, 2013.

[15] A. Xu, A. Kalmbach, and G. Dudek. Adaptive Parameter EXploration (APEX): Adaptation of robot autonomy from human participation. In *Proc. of the IEEE Int. Conf. on Rob. and Auto. (ICRA'14)*, 2014.