2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Getting to Know One Another: Calibrating Intent, Capabilities and Trust for Human-Robot Collaboration

Joshua Lee, Jeffrey Fong, Bing Cai Kok, and Harold Soh
Dept. of Computer Science, National University of Singapore
joshua_lks@u.nus.edu, {jfong, kokbc, harold}@comp.nus.edu.sg

***Abstract*— Common experience suggests that agents who know each other well are better able to work together. In this work, we address the problem of calibrating intention and capabilities in human-robot collaboration. In particular, we focus on scenarios where the robot is attempting to assist a human who is unable to directly communicate her intent. Moreover, both agents may have differing capabilities that are unknown to one another. We adopt a decision-theoretic approach and propose the TICC-POMDP for modeling this setting, with an associated online solver. Experiments show our approach leads to better team performance both in simulation and in a real-world study with human subjects.**

I. INTRODUCTION

As robots are increasingly deployed into our homes and workplaces, it is crucial that human users trust their robot collaborators appropriately. In particular, it is important to mitigate under-trust and over-trust in robots, which can lead to unsatisfactory outcomes [1], [2]. Building upon prior work that shows human trust in robots is dependent on the robot's intention (or policy) and capabilities [3], [4], we focus on *calibrating* these two constructs for trust-based human-robot collaboration.

We consider assistive scenarios where the robot is helping the human to accomplish a particular goal, but is unaware of the human's intent. As such, the robot has to learn the goal through interaction. Unlike a majority of existing work, we address the case where the human and robot have asymmetric capabilities that are unknown to one another. This setting is important as non-expert users may be unaware of the robot's programming and physical capabilities. Poorly calibrated capability models can damage human users' perception of the robot's trustworthiness [5], which in turn affects their willingness to cooperate with it. Likewise, humans have differing proficiency at tasks, and failure to recognize capability differences may lead to incorrect inferences about a person's goals. In short, failure to acknowledge each other's capabilities can impede collective performance.

We undertake a decision theoretic approach to the problem and contribute the Trust-Intent-Capability-Calibration POMDP (TICC-POMDP). The key idea underlying our approach is a structured transition, which enables us to incorporate both robot and human capability models as parameters in an expanded state-space model. As estimations of capability are likely to persist over multiple rounds of interactions (and even tasks), we encourage calibration via an additional calibration reward that incentivizes the robot to engage in teaching behaviors that inform the human about its (in)capabilities. To solve the TICC-POMDP, we propose an online planning Monte-Carlo method, the TICC-MCP, which scales to large state-spaces. Code and an appendix are available at https://github.com/clear-nus/TICC-MCP.

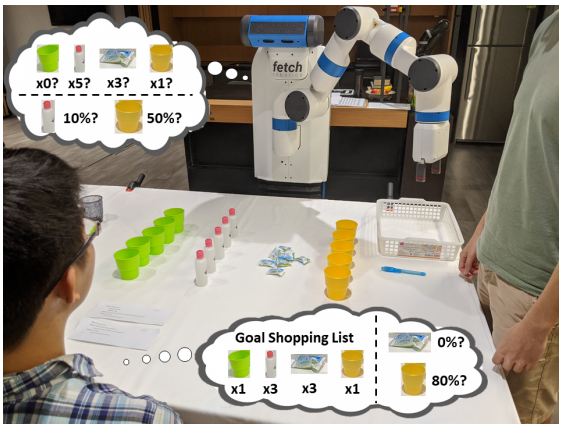


Fig. 1. Calibration experiment with human subjects. The human and the Fetch robot have to work together to complete a shopping task. However, only the human knows the goal shopping list, and is unable to communicate it directly to the robot. Moreover, both agents have imperfect capabilities and are unable to pick up certain items. Using our TICC-POMDP (and associated TICC-MCP solver), the agents undergo intent-capability calibration and update their beliefs of each others' capabilities over time. Experiments show that our approach leads to more accurate beliefs over intention and capabilities, and in-turn, higher task rewards and trust.

We present both simulation and real-world experiments with human subjects ($n = 28$). Our simulation experiments focused on the algorithmic properties of the TICC-MCP and we find our method outperforms the standard POMCP (which lacks capability models). Our human subject experiments (Fig 1) show that our approach earned higher rewards with actual humans and we find evidence that suggests our method induced higher levels of trust in the robot.

To summarize, our key contributions are:

- A novel decision-theoretic formulation for calibrating intent and capability in human-robot collaboration;
- The TICC-MCP online planning algorithm for (approximately) solving the TICC-POMDP;
- Simulation and real-world human subject experimental findings showing that intent and capability calibration is beneficial over repeated interactions.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

II. BACKGROUND

In this section, we briefly cover the necessary background material; we first give an overview of sequential decision-making frameworks, specifically the Markov decision process (MDP) and its generalization, the partially-observable MDP (POMDP) and Bayes Adaptive POMDP (BA-POMDP). Then, we discuss inverse reinforcement learning, and trust in human-robot collaboration.

A. Decision-Making Frameworks

Markov Decision Processes. A MDP models a sequential decision-making problem as a tuple $\langle S, A, T, R, \gamma \rangle$, where S is the set of environmental states, A is the set of actions available to the agent, and the transition function $T(s_t, a_t, s_{t+1})$ models the probability of transitioning to state $s_{t+1} \in S$ after taking action $a_t \in A$ in state s_t . At each state, the agent receives a reward $R(s)$[1] and the agent's goal is to obtain the optimal policy $\pi^* : S \rightarrow A$ that maximizes the expected return: $\pi^* = \arg\max_\pi \mathbb{E}\left[\sum_t \gamma^t R(s_t)\right]$ where $\gamma \in [0, 1]$ is a discount factor and the expectation is over the trajectories generated by following π.

Partially-Observable MDPs. In contrast to MDPs where states are fully-observable, POMDPs model sequential decision-making *under uncertainty* where agents can only make observations that arise from the underlying state. We augment the MDP with two additional elements: the set of observations Z and the observation function $O(s_{t+1}, a_t, z_t) = p(z_t | s_{t+1}, a_t)$, which specifies the probability of observing $z_t \in Z$ after having taken action $a_t \in A$ and arriving in state $s_{t+1} \in S$. Because the agent cannot observe the state directly, it has to rely on its interaction history $h_t = \langle a_o, z_1, \ldots, a_{t-1}, z_t \rangle$ to infer a belief $b(s)$ over states. The policy $\pi(b)$ thus maps beliefs onto actions. As in the MDP case, the agent aims to maximize its expected return given an initial belief b_0:

$$\pi^* = \arg\max_\pi V^\pi(b_0) = \arg\max_\pi \mathbb{E}\left[\sum_t \gamma^t R(s_t) \,\middle|\, b_0\right] \quad (1)$$

where $V^\pi(b_0)$ is the value function and the expectation is over trajectories obtained by following π starting from belief b_0. In general, solving POMDPs is intractable but efficient and scalable methods exist for approximate planning [6], [7], [8], [9].

Bayes-Adaptive POMDP. The environment model is often not perfectly known to the agent. The Bayes-Adaptive POMDP (BA-POMDP) [10] is an example of the Bayesian Reinforcement Learning approach for learning POMDP transition and observation models. The BA-POMDP uses Dirichlet distributions to represent uncertainty over the transition and observation models; the agent maintains a vector χ with the experience counts of $\langle s, a, s', z \rangle$. The agent is uncertain about the true count vector and this uncertainty is represented as hidden state in POMDP, i.e., the state space is augmented with the Dirichlet parameters.

As BA-POMDPs have infinite state space, Ross *et al.* [10] proposed a reduction to a finite model, and developed an online lookahead planner. Extending this work, Katt *et al.* [11] developed the BA-POMCP algorithm, which uses POMCP [12] to solve BA-POMDPs more efficiently via sample-based online planning. Our collaborative model TICC-POMDP is a BA-POMDP with specific structure, and we propose a variant of BA-POMCP algorithm.

B. Inferring Reward Functions

In this paper, we cast intent inference as problem of inverse reinforcement learning (IRL), i.e., recovering the reward function of the agent given its policy or observed behavior [13]. More formally, let $\mathcal{M}_{\setminus R}$ be an MDP without the reward function R of an expert. IRL seeks a reward function $\hat{R}_E$ that best describes a set of observed expert demonstrations $D = \{\tau_i\}_i^N$. The approach underlying many IRL methods (e.g., [14], [15], [16]) is to update (or optimize) a parametric reward function $R(s; \theta)$ until it is consistent with the observed behavior, under constraints or heuristics to address ambiguity [13], [15], [17].

One potential drawback of the above methods is the underlying assumption that the demonstrator acts to maximize expected returns. However, expert demonstrations may be suboptimal due to cognitive biases in planning [18], [19], [20], risk sensitivity [21], imperfect beliefs about real world dynamics [22] and pedagogy behavior [23]. Here, we consider situations where the expert demonstrations are "suboptimal" in that the robot and human may have incorrect models of each other's capabilities.

C. Trust in Human Robot Collaboration

Trust is important in shaping how humans interact with one another [24]. As with humans, the issue of human trust in automation has also been identified as a key determinant to successful human-robot collaboration [4], [1]. For instance, trust in automation has been shown to increase performance in human-robot teams [25], affect reliance on automation [26] and determine whether or not humans will choose to retaliate against an adversarial robot [27].

There exist various definitions of trust, ranging from a belief [28] to an operator in a logic [29]. We adopt a formal definition of trust in the human-robot collaboration literature: trust is a latent variable that summarizes a human's interactions with the robot and determines whether or not one would be willing to rely on the autonomous robot [30], [31].

Through this formalization, trust in automation can also be exploited to further improve both planning [30] and predictive algorithms that can capture how human trust in a robot transfers across multiple tasks [32]. Most pertinently, the explicit incorporation of human factors, such as the human's preference for explainability [33], into the design of autonomous agents could implicitly foster greater trust in autonomous robotic agents. In line with this, we explore the hypothesis that calibrated human and robot capability models can encourage humans to develop trust in robots.

[1]The reward may also depend on the next state s_{t+1} and action a_t.

III. Calibrating Intent and Capabilities with TICC-POMDP

In this section, we describe the TICC-POMDP for calibrating both intention and capabilities. We consider planning from the *perspective of the robot* and at a high level, our formulation is a POMDP augmented with human and robot capability models. Specifically, we model the robot's belief over the human's capabilities and the human's belief over the robot's capabilities. These models enter into a *structured* transition function. Overall, the TICC-POMDP can be seen as a BA-POMDP with latent reward and capability parameters, and a reward calibration component.

The following elaborates upon the above; we first describe the key ideas underlying our approach, and then apply them to the human-robot collaboration setting.

A. Action Success & Failures and Factorized Transitions

In our model, actions can either succeed or fail; for example, an agent may accidentally drop an object when trying to pick it up; the agent did not *intend* to drop the item, but may do so due to its (lack of) dexterity at handling items, i.e., its *capability*. In other words, we consider the capability of agents to be the probability that actions taken are successful. In typical POMDPs, action success/failure is typically *implicit* within the model's transition probabilities. The key idea in our approach is to make these "action outcomes" *explicit*. For expositional simplicity, we discuss binary success/failure action outcomes but our formulation generalizes to different levels of success/failure.

More precisely, for every action a, we define a successful action a^{S} and a failed action a^{F}. We can then model state transitions as a marginalization of a factorized transition probability over *action outcomes* $\tilde{a}$:

$$\begin{aligned} P(s_{t+1}|s_t, a_t) &= \sum_{\tilde{a}_t \in \{a_t^{\mathrm{S}}, a_t^{\mathrm{F}}\}} P(s_{t+1}|s_t, \tilde{a}_t, a_t) P(\tilde{a}_t|s_t, a_t) \\ &= \sum_{\tilde{a}_t \in \{a_t^{\mathrm{S}}, a_t^{\mathrm{F}}\}} P(s_{t+1}|s_t, \tilde{a}_t) P(\tilde{a}_t|a_t) \end{aligned} \tag{2}$$

where we have assumed that (i) given the action outcome $\tilde{a}_t$, the next state is independent of the action a_t, and (ii) given the action a_t, the action outcome $\tilde{a}_t$ is independent of the state s_t. The latter conditional independence assumption is not strictly necessary but reduces model complexity.

Intuitively, the structured transition comprises two parts:

$$\underbrace{P(s_{t+1}|s_t, \tilde{a}_t)}_{\text{World Physics}} \quad \underbrace{P(\tilde{a}_t|a_t)}_{\text{Agent Capability}} \tag{3}$$

Decomposing the transition probability in this manner has benefits; it represents the world physics separately from the agent's capabilities, which can ease data requirements if the transition model is learnt from data—particularly if one part (e.g., the world physics) is known. Moreover, different agent models can be substituted into the POMDP without having to re-learn or re-model the entire transition dynamics. Likewise, the same agent model can be applied to different environments (with the same action set).

In this work, we assume that the world physics is known to both agents, but the individual capabilities are not mutually-known, i.e., agents know their own abilities, but are unaware of the capabilities of other agents.

B. Modeling Uncertainty over Capabilities

Since capabilities are unknown, our POMDP is not fully-specified. We overcome this issue by modeling the *uncertainty* over potential agent capabilities. We adopt the BA-POMDP approach, which enables our robot to consider this uncertainty during *planning*, and *learn* the capabilities from interaction experience.

Given binary outcomes, the agent capabilities $P(\tilde{a}|a)$ can be modeled as Bernoulli distributions (or more generally, categorical distributions for K discrete action outcomes) with unknown parameters $\mathbf{p}$. For each capability distribution, we apply a Dirichlet prior with parameters $\boldsymbol{\chi} \in \mathcal{X}$ and thus,

$$\begin{aligned} P(\tilde{a}|a; \boldsymbol{\chi}) &= \mathbb{E}_{\mathbf{p} \sim \mathrm{Dir}(\boldsymbol{\chi}_a)}[P(\tilde{a}|a, \mathbf{p})] \\ &= \frac{\boldsymbol{\chi}_a^{\tilde{a}}}{\sum_{\tilde{a}'} \boldsymbol{\chi}_a^{\tilde{a}'}} \end{aligned} \tag{4}$$

where $\boldsymbol{\chi}_a^{\tilde{a}}$ is an "experience count" representing the number of times $\tilde{a}$ occurs when action a is taken. This expectation is computationally efficient and simple to compute.

As with general BA-POMDPs, the parameters $\boldsymbol{\chi}$ are embedded into the state. We construct a new POMDP with augmented state set $\hat{S} = S \times \mathcal{X}$. The transition probabilities for the original state s follow Eqn. (2) above, and $\boldsymbol{\chi}$ is deterministically updated via,

$$\boldsymbol{\chi}_{t+1} = \boldsymbol{\chi}_t + \Delta_a^{\tilde{a}} \tag{5}$$

where $\Delta_a^{\tilde{a}}$ is simply a one hot vector with value 1 at the position corresponding to the $\langle a, \tilde{a} \rangle$. As long as the robot can observe the action and action outcome (or some function of these two elements), it can update the capability model.

C. Calibrating Intent and Capabilities

Let us now apply the ideas developed above to human-robot collaboration. We denote the human and robot actions as a^{H} and a^{R}, respectively. Given both actions, the world state transitions according to $P(s_{t+1}|s_t, a^{\mathrm{H}}, a^{\mathrm{R}})$. Recall that we seek a policy for our robot whose objective is to help the human accomplish her goal. However, the robot is uncertain about both the intent and capabilities of its human partner. Moreover, the human may be uncertain about the robot's capabilities.

Calibrating Intention. We assume that the POMDP reward function is parameterized by $\theta \in \Theta$ where Θ is a finite set of possible rewards. In our scenario, the intention of the human (represented by θ) is unknown to the robot. Similar to other POMDP-based human-robot collaboration models [34], [35], a straightforward way to enable the robot to learn θ is to embed it into the state—we augment the POMDP's state with θ. Although θ is latent, human behavior is informative about its value; the human is assumed to act consistently to achieve her objectives. In this work, both human actions

(and the corresponding action outcomes) are observable by the robot, and the human follows a fixed policy—this last assumption avoids recursive planning and future work may look into alternative approximations (e.g., [36]). Regardless, learning the human's intention without understanding the human's capability may lead to incorrect estimates.

Calibrating Human and Robot Capabilities. Let $\boldsymbol{\psi} \in \Psi$ be a set of Dirichlet parameters representing the human's capability. The robot maintains a belief over the human's capability by incorporating $\boldsymbol{\psi}$ into the state representation. During state transitions, $\boldsymbol{\psi}$ is updated similar to Eqn. (5).

We also calibrate the human's belief over the robot's capability, i.e., we model the human's belief over the robot's capability, and compare it against the robot's true capabilities. We assume that the human is Bayes rational and represent the human's belief over the robot's capability as another set of Dirichlet parameters $\boldsymbol{\phi} \in \Phi$. Similar to the human capability parameters $\boldsymbol{\psi}$, the robot capability belief parameters $\boldsymbol{\phi}$ are also incorporated into the state, with the transition probabilities adjusted in a similar fashion.

Putting all these components together results in an augmented state space $\bar{S} = S \times \Theta \times \Psi \times \Phi$. The transition probability is given by,

$$\bar{T}(\bar{s}_t, a_t^{\mathrm{R}}, \bar{s}_{t+1}) = \sum_{a^{\mathrm{H}}} P(\bar{s}_{t+1}|\bar{s}_t, a_t^{\mathrm{R}}, a^{\mathrm{H}}) P(a^{\mathrm{H}}|\bar{s}_t, a_t^{\mathrm{R}})$$

$$= \sum_{\tilde{a}^{\mathrm{R}}} \sum_{\tilde{a}^{\mathrm{H}}} \sum_{a^{\mathrm{H}}} \Bigg[P(s_{t+1}|s_t, \tilde{a}^{\mathrm{H}}, \tilde{a}^{\mathrm{R}}) P^*(\tilde{a}^{\mathrm{R}}|a_t^{\mathrm{R}}) P(\tilde{a}^{\mathrm{H}}|a^{\mathrm{H}}; \psi_t)$$

$$P(a^{\mathrm{H}}|\bar{s}_t, a_t^{\mathrm{R}}) \delta_{\theta_{t+1}, \theta_t} \delta_{\psi_{t+1}, \psi_t + \Delta_{a^{\mathrm{H}}}^{\tilde{a}^{\mathrm{H}}}} \delta_{\phi_{t+1}, \phi_t + \Delta_{a^{\mathrm{R}}}^{\tilde{a}^{\mathrm{R}}}} \Bigg] \quad (6)$$

where $\delta_{i,j}$ is the Kronecker's delta function, which is 1 if $i = j$, and 0 otherwise. Note that unlike the capability parameters, the intention parameter θ does not change; a different objective is chosen at the beginning of episode and remains fixed until the next episode.

Teaching Actions. In order to better facilitate pedagogy behaviour, we expand our action set with failure actions $\hat{a}^{\mathrm{F}}$, and allow both the human and robot to deliberately choose failure actions when they wished to indicate (in)capability. Performing $\hat{a}^{\mathrm{F}}$ will always lead to the outcome action a^{F}. Note that in practice, it may be desirable for the robot to avoid actual failures; rather, choosing $\hat{a}^{\mathrm{F}}$ would cause the robot to perform a behavior that is indicative of its incapability [37]. As such, $\hat{a}^{\mathrm{F}}$ can be regarded as a communicative action.

Task and Calibration Rewards. Apart from its primary objective of achieving the task goal, we incentivize the robot to calibrate $\boldsymbol{\phi}$ to its *actual* capabilities. This is achieved via an additional *calibration reward* at each time-step. This calibration reward is computed based on the similarity between $P(\tilde{a}^R|a^{\mathrm{R}}; \phi)$ and the true capabilities $P^*(\tilde{a}^R|a^{\mathrm{R}})$. We tested several similarity measures and found a simple area overlap (AO) to work well; given two categorical distribution parameters $\mathbf{p}$ and $\mathbf{q}$, we sum the minimum of the corresponding components of the two distributions, i.e., $\sum_i \min(p_i, q_i)$. We hypothesized the calibration reward would lead to more teaching behaviors, i.e., the robot would attempt to teach the human about its (in)capabilities. This could enable better calibration, which would in-turn lead to better team performance over *multiple* episodes or interaction rounds.

D. Solver Design: TICC-MCP

In order to solve the TICC-POMDP, we developed an online planning approach we call TICC-MCP (Algorithm 1 and Fig. 1 in the Online Appendix). Online planning methods have been shown to scale well to large state spaces, and the incorporated models allow us to exploit structure and prior knowledge. Our algorithm is similar to the BA-POMCP, with the following two differences:

Keeping Additional Statistics. In order to simulate a^{H}, we need to compute $P(a_t^{\mathrm{H}}|\bar{s}_t, a_t^{\mathrm{R}})$ in the search tree. Given the history h at the parent node and the history $ha^{\mathrm{R}}a^{\mathrm{H}}$ at the current node, we keep an additional vector of statistics $V_\theta(ha^{\mathrm{R}}a^{\mathrm{H}})$ (at each node), which represents the Q-values for the human under different θ. Additionally, we keep $N_\theta(ha^{\mathrm{R}}a^{\mathrm{H}})$ and $N_\theta(ha^{\mathrm{R}})$, which represent the number of times $ha^{\mathrm{R}}a^{\mathrm{H}}$ and ha^{R} is visited for each θ. These are used to compute the Upper Confidence Bound (UCB1) [38] values for sampling the human actions. The backup rule for $V_\theta(ha^{\mathrm{R}}a^{\mathrm{H}})$, $N_\theta(ha^{\mathrm{R}}a^{\mathrm{H}})$ and $N_\theta(ha^{\mathrm{R}})$ is akin to that for $V(ha^{\mathrm{R}})$ and $N(ha^{\mathrm{R}})$.

Sampling Human and Robot Actions. At each node, probability of action a^{H} is computed using UCB1 [39]:

$$P(a^{\mathrm{H}}) = P\left(a^{\mathrm{H}} \middle| V_\theta(ha^{\mathrm{R}}a^{\mathrm{H}}) + c\sqrt{\frac{\log N_\theta(ha^{\mathrm{R}})}{N_\theta(ha^{\mathrm{R}}a^{\mathrm{H}})}}\right) \quad (7)$$

and the action outcome $\tilde{a}^{\mathrm{H}}$ is sampled from the human capability model, $\tilde{a}^{\mathrm{H}} \sim \sum_{a^{\mathrm{H}}} P(\tilde{a}^{\mathrm{H}}|a^{\mathrm{H}}; \psi) P(a^{\mathrm{H}})$. Robot actions are similarly sampled with the appropriate histories, and action outcomes from $P^*(\tilde{a}^{\mathrm{R}}|a^{\mathrm{R}})$.

E. Comparison to Related Work

The TICC-POMDP (and associated solver) is related to a body of literature applying POMDPs towards human robot collaboration (HRC). For example, recent work [1], [30] has developed trust-based POMDPs that enable robots to consider the human's underlying trust in the robot during planning. Other works have contributed technical innovations for dealing with uncertainty and the complexity of HRC POMDPs, e.g., a Bayesian method to learn the state space and estimate the observation/transition functions [34], and hierarchical structure to reduce planning complexity [40]. Similar to our solution technique, [35] modified the vanilla POMCP for large observation spaces associated with HRC POMDPs. Note that the above works do not consider *both* human and robot capabilities and their relationship to intent inference, which is a unique aspect of this paper.

The TICC-POMDP is similar to Cooperative Inverse Reinforcement Learning (CIRL) [23] in that both TICC-POMDP and CIRL assume a collaborative setting where the human and robot share the same reward function whose parameters is only known to the human. However, behaviour in CIRL is defined by a joint robot-human policy, while behaviour in TICC-POMDP is defined solely by robot policy with an internal human model. TICC-POMDP relaxes the assumption on human rationality (also considered in [36]) and incorporates explicit capability models.

Our work is also related to methods that model robot (in)capability. For example, recent work has proposed conceptual HRC models based on Theory of Mind [41], planning with robot capability models [42], and a game-theoretic approach for teaching humans about robot capabilities [43]. Very little work has looked into incorporating human capability models into HRC; a possible exception [44] modelled human capabilities with Bayesian Networks for planning.

IV. Simulation Experiments

We evaluated our model using both simulated and real-world experiments. Our main hypothesis was that calibrating a human capability model resulted in higher task rewards over the long term, i.e., over multiple episodes or interaction rounds. We compared the TICC-MCP (TICC) against a standard POMCP algorithm (STD). The standard POMCP performed intent inference but did not explicitly model capability nor had a reward calibration component.

This section describes our simulation experiments where we focused on evaluating the algorithmic aspects of our approach. We were interested in whether our method was able to infer the intent and capabilities of a simulated human, under various parameter settings.

Domain. Our task was in the shopping domain where the human and robot collaborate to collect a bag of items within a limited horizon. There are n item types and m possible shopping lists, and the world state is a n-tuple representing the quantity of each item already collected. Although seemingly abstract and small, the problem described above is challenging and similar to domains explored in recent work [36]. The largest of our experiments involved more than 10^{10} world states.

The human and robot would collaborate over multiple episodes (rounds). At the beginning of each round, the human observes the intended shopping list, which corresponds to a specific reward parameter θ. Only the human knows the shopping list at the outset, while the robot starts with a uniform belief over each list. Both agents are rewarded based on how closely their item bag matches the goal list. At each time-step, the robot and human each chooses to pick-up an item, or does nothing. Each item type can be chosen multiple times throughout each episode (there is an unlimited number of items). Neither the human nor the robot are perfectly capable and may fail to pick up certain items.

Manipulated Variables. To evaluate our method, we varied the number of search samples, shopping item types (i.e., the size of action and observation spaces), and shopping lists (the size of the reward space).

Dependent Measures. We evaluated the performance of the two algorithms using the the task reward received and the correctness of the belief over the human's intent θ and human's capability ψ. The correctness of the intent was measured by average likelihood of the true θ given the model. Capability correctness was measured as the area of overlap between the normalized belief distribution and the true capability distribution. The measures were computed over 50 independent runs.

Method. The robot was initialized to believe that the human is fully capable. Each experiment consisted of two stages: a learning stage and an evaluation stage. In the learning stage, we first ran 5 rounds of simulation for the robot to learn about the human's capability. In the evaluation stage, we ran another 5 rounds, from which the average task reward was calculated. Note that the human capability parameters ψ carry over the rounds, but θ is initialized randomly at the beginning of each round.

Results. In brief, the simulation results indicate that TICC-MCP agent was better able to infer the human's intent and capability, which led to improved task rewards over the long-term. We see in Fig. 2 that our approach consistently achieved higher average task rewards across the variable changes. Fig. 3 and Fig. 4 show an illustrative result with 10 shopping lists, 5 items, and 50k search samples. TICC was able to quickly learn the human's capabilities. It was better able to estimate the correct θ leading to higher performance compared to STD, particularly in the evaluation stage.

V. Human Subject Experiments

This section describes our human subject experiments where we sought to assess the TICC-MCP. Our primary hypotheses were that:

- **H1** Calibrated intent and capability models improves performance (i.e. task rewards) over a sequence of episodes.
- **H2** Mutual calibration induces higher trust in the assistive robot.

Experimental Setup. Participants interacted with a real Fetch robot in the collaborative item-shopping task in a table-top setting (see Fig. 1). Four types of items were available for purchase: yellow cups, sweets, corrosive cleaning agents and green cups. Half of the yellow cups were "faulty" (marked with a cross) while the other half are "good" (marked with a tick); note that the markings were only visible under UV light. This simulates a grocery shopping scenario where it may be difficult to distinguish between good and faulty products with the naked eye (e.g., good or bad fruits may be hard to distinguish). In our experiment, the human is unable to pick up the corrosive cleaning agent, and the robot is unable to pick up the small sweets. The human has a 50% chance of picking a "good" yellow cup since half of them are faulty. In contrast, the robot has an 80% probability of picking a "good" yellow cup.

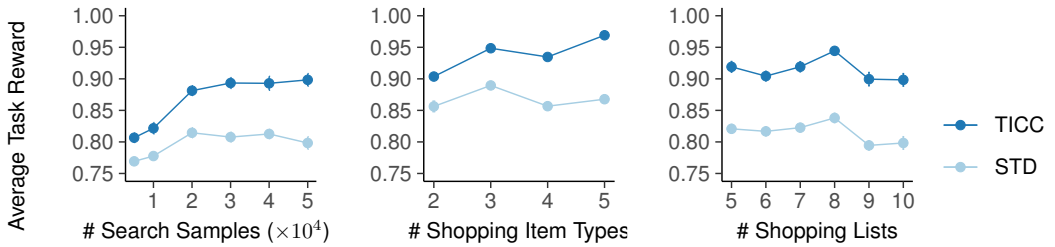


Fig. 2. The average task rewards obtained by TICC-MCP and standard POMCP during evaluation stage with varying number of search samples (left), shopping item types (middle) and shopping lists (right). Error bars indicate one standard error.

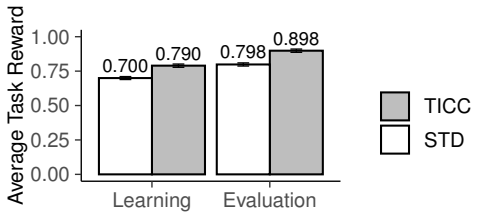


Fig. 3. The average task rewards over simulation rounds. Error bars indicate one standard error.

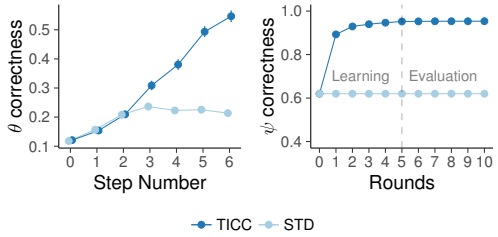
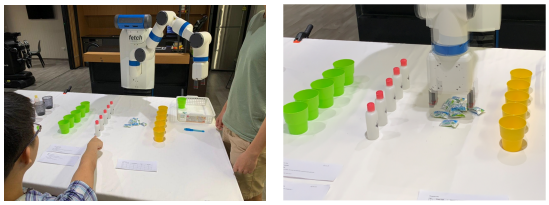


Fig. 4. θ belief correctness over number of steps taken in a single episode (left). ψ belief correctness of TICC-MCP and standard POMCP over simulation rounds (right). Error bars indicate one standard error.

The horizon was fixed at 6 steps for both agents. At each step, the robot will make the first move followed by the human. Each of the shopping lists had either 8 or 9 items. As such, the human participant is unable to single-handedly fulfil the shopping list due to her imperfect capability and the limited horizon. Each experiment lasted ≈ 60 minutes.

Procedure. Participants entered the lab and were briefed about the task. They were also told that they have two types of privileged information: 1) their own capability and 2) the actual shopping list for that round. They then engaged in a practice round where a human demonstrator took the place of the Fetch robot. This was done to avoid carryover effects from the practice to the main experiment.

Thereafter, they performed 5 consecutive rounds of the task with the real robot. A different shopping list, selected by a seeded random number generator, was used for each round. The order of the shopping list was fixed for all participants to reduce experimental variability and to allow for a better comparison between the two algorithms. Participants received \$1 or \$0.50 bonus if they completed the shopping list or if the number of items exceeded what was required respectively. On every turn, both the participant and robot can choose to either pick an item or do nothing. Additionally, participants can choose to indicate to the robot that they want to pick



(a) (b)

Fig. 5. Examples of communicative actions. (a) The human indicates to the robot that he wants to pick the corrosive cleaning agent, but is unable to; (b) The robot executes a communicative action with the participant about its capability by hovering above the sweets.

an item, but is unable to (Fig. 5a). Likewise, the robot may choose actions to communicate with the participant (Fig. 5b).

Participant Assignment. A total of 28 participants (mean age = 22.8, 16 females) were recruited from the university community. The experiment was designed to be between-subjects with 14 participants in each condition. They were randomly assigned to play with either the TICC-MCP or standard POMCP robot.

Dependent Measures. We used both objective and subjective measures to evaluate the performance of the algorithms. There were three objective performance measures:

- The average task reward for the final three rounds
- The calibration score of belief of robot's capability ϕ
- The calibration score of belief of human's capability ψ

We also collected the subjective measures shown in Table I throughout the course of the experiment.

Results. The data from 4 participants were removed due to a consistency check question failure[2]. All analyses were carried out with available-case analysis. As seen in Fig. 6, the mean calibration scores for both ψ and ϕ show an increasing trend. The increase in ψ calibration scores suggests that the TICC-MCP robot can effectively learn about the participants' capabilities. The increase in ϕ calibration scores suggests that the TICC-MCP robot influenced the participants' belief of its capability. More details can be seen in Fig. 7 which indicates the participants beliefs over the robot's capabilities gradually matched reality.

In brief, we found that the TICC-MCP algorithm led to significantly higher average task rewards in the final three

[2]They responded that the robot was worse at picking-up the bottle at the last round, despite the fact that they cannot pick up the bottle.

TABLE I

SUBJECTIVE MEASURES IN THE HUMAN EXPERIMENT.

	Subjective Measures
Before Interaction	**General Perception of Robot** - Negative Attitudes toward Situations and Interactions with Robots [45] - Schaefer's Trust Perception Scale [46]
Before 1^{st} Round	**Perceived Relative Capability** *(Robot is better/worse/both are equally capable)* - Do you think the robot is better than you at picking the [insert item name]?
After each Round	**Perceived Relative Capability** *(Robot is better/worse/both are equally capable)* - Do you think the robot is better than you at picking the [insert item name]? **Human-Robot Trust** *(5-point Likert Scale)* - I trust the robot to collaborate with me on this task - The robot trusts me to collaborate with it on this task **Human's First-Order Belief of Robot's Beliefs** *(5-point Likert Scale)* - The robot understands what I am capable of - The robot is able to infer my intentions
After Interaction	**General Perception of Robot** - Negative Attitudes toward Situations and Interactions with Robots - Schaefer's Trust Perception Scale

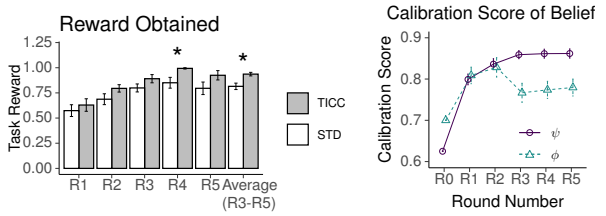


Fig. 6. Task rewards over rounds (left). Convergence of belief over both ψ and ϕ across simulation rounds for TICC-MCP agent (right) . Error bars indicate one standard error. Plots marked with * indicate statistical significance.

rounds (Fig. 6, $t(22) = 3.474, p = 0.00215$). Separate two-sample t-tests at each of the final three rounds suggest that participants who worked with the TICC-MCP robot achieved higher rewards (i.e. the t-values are all positive) (Round 3: $t(22) = 1.599, p = 0.124$; Round 4: $t(22) = 2.881, p = 0.0087$; Round 5: $t(22) = 1.669, p = 0.103$), although only the difference in Round 4 achieved statistical significance when $\alpha = 0.05$. The difference between the averaged rewards and the reward at Round 4 is statistically significant after adjusting for multiple comparisons using Bonferroni correction (adjusted-$\alpha = 0.0125$). These results support hypothesis **H1**.

We carried out planned two-sample comparisons for each of the four survey questions at the end of the final round. Participants in the TICC-MCP condition reported having higher levels of trust in the robot ($t(21) = 2.224, p = 0.0372$) (see Fig. 8 (top left)). This provides evidence for

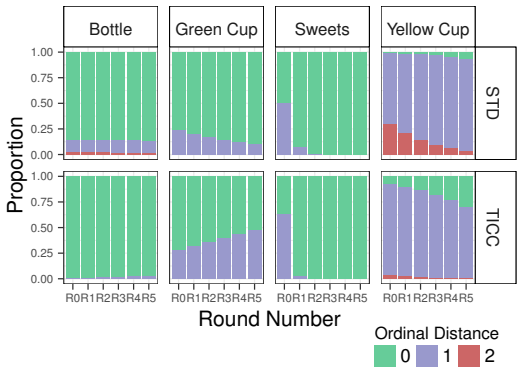


Fig. 7. Predicted probabilities (based on fixed-effects estimates) of the ordinal distance at each time point. An ordinal distance of 0 indicates that the participant's beliefs of the robot's capability agrees with reality.

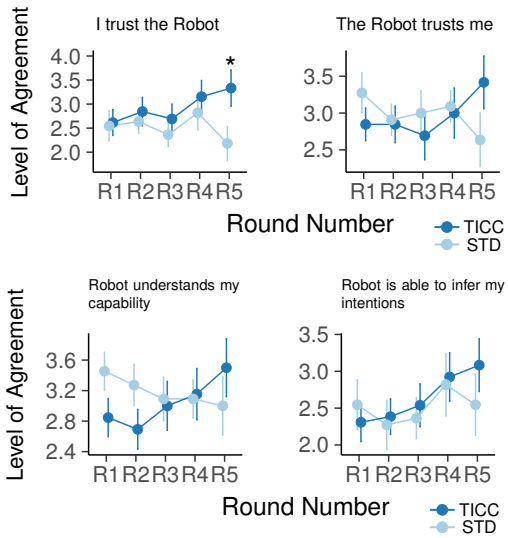


Fig. 8. Subjective measures of Human-Robot Trust and the participants' perception of the robot's ability to understand their capabilities and intentions. Error bars indicate one standard error. Plots marked with * indicate statistical significance.

hypothesis **H2**. The differences for other three questions did not achieve statistical significance when $\alpha = 0.05$. Nevertheless, Fig. 8 suggests that participants in different conditions perceived the robot differently as they interacted with it.

VI. CONCLUSION

This paper contributes the TICC-POMDP and associated TICC-MCP online solver, which our experiments show can enable better human-robot cooperation over multiple interactions. Calibrating mutual beliefs of intention and capability resulted in both higher team performance and higher human trust. As future work, we intend to explore applying TICC to longer-term real-world scenarios and with other salient aspects, e.g., across multiple tasks [32], [31] or where attention is crucial [47].

ACKNOWLEDGEMENTS

This work was supported by the SERC, A*STAR, Singapore, through the National Robotics Program under Grant No. W1925d0054.

REFERENCES

[1] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Trust-aware decision making for human-robot collaboration: Model learning and planning," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 2, pp. 1–23, 2020.

[2] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 101–108.

[3] Y. Xie, I. P. Bodala, D. C. Ong, D. Hsu, and H. Soh, "Robot capability and intention in trust-based decisions across tasks," in *2019 14th ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 39–47.

[4] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[5] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *2010 5th ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 203–210.

[6] H. Kurniawati, D. Hsu, and W. S. Lee, "Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces." in *Robotics: Science and Systems*, 2008.

[7] A. Somani, N. Ye, D. Hsu, and W. S. Lee, "Despot: Online pomdp planning with regularization," in *Advances in Neural Information Processing Systems*, 2013, pp. 1772–1780.

[8] P. Karkus, D. Hsu, and W. S. Lee, "Qmdp-net: Deep learning for planning under partial observability," in *Advances in Neural Information Processing Systems*, 2017, pp. 4694–4704.

[9] J. Pineau, G. Gordon, S. Thrun, *et al.*, "Point-based value iteration: An anytime algorithm for pomdps," in *18th Intl. Joint Conf. on Artificial Intelligence*, 2003, pp. 1025–1032.

[10] S. Ross, B. Chaib-draa, and J. Pineau, "Bayes-adaptive pomdps," in *Advances in Neural Information Processing Systems*, 2008, pp. 1225–1232.

[11] S. Katt, F. A. Oliehoek, and C. Amato, "Learning in pomdps with monte carlo tree search," in *34th Intl. Conf. on Machine Learning*, 2017, pp. 1819–1827.

[12] D. Silver and J. Veness, "Monte-carlo planning in large pomdps," in *Advances in Neural Information Processing Systems*, 2010, pp. 2164–2172.

[13] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *arXiv preprint arXiv:1806.06877*, 2018.

[14] S. J. Russell, "Learning agents for uncertain environments," in *11th Conf. on Computational Learning Theory*, 1998, pp. 101–103.

[15] A. Y. Ng, S. J. Russell, *et al.*, "Algorithms for inverse reinforcement learning." in *17th Intl. Conf. on Machine Learning*, 2000, pp. 663–670.

[16] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *21st intl. conf. on Machine learning*, 2004.

[17] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *AAAI*, 2008, pp. 1433–1438.

[18] L. Bergen, O. Evans, and J. Tenenbaum, "Learning structured preferences," in *Annual Meeting of the Cognitive Science Society*, vol. 32, no. 32, 2010.

[19] O. Evans and N. D. Goodman, "Learning the preferences of bounded agents," in *NIPS Workshop on Bounded Optimality*, vol. 6, 2015.

[20] O. Evans, A. Stuhlmüller, and N. Goodman, "Learning the preferences of ignorant, inconsistent agents," in *Thirtieth AAAI Conf. on Artificial Intelligence*, 2016.

[21] A. Majumdar, S. Singh, A. Mandlekar, and M. Pavone, "Risk-sensitive inverse reinforcement learning via coherent risk models." in *Robotics: Science and Systems*, 2017.

[22] S. Reddy, A. Dragan, and S. Levine, "Where do you think you're going?: Inferring beliefs about dynamics from behavior," in *Advances in Neural Information Processing Systems*, 2018, pp. 1454–1465.

[23] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3909–3917.

[24] D. Balliet and P. A. Van Lange, "Trust, conflict, and cooperation: a meta-analysis." *Psychological Bulletin*, vol. 139, no. 5, p. 1090, 2013.

[25] S. You and L. Robert, "Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction," in *52th Hawaii Intl. Conf. on System Sciences*, 2018, pp. 8–11.

[26] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Intl. Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[27] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, ""I Don't Believe You": Investigating the effects of robot trust violation and repair," in *2019 14th ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 57–65.

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

[29] X. Huang, M. Kwiatkowska, and M. Olejnik, "Reasoning about cognitive trust in stochastic multiagent systems," *ACM Transactions on Computational Logic (TOCL)*, vol. 20, no. 4, pp. 1–64, 2019.

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

[31] H. Soh, Y. Xie, M. Chen, and D. Hsu, "Multi-task trust transfer for human–robot interaction," *The Intl. Journal of Robotics Research*, vol. 39, no. 2-3, pp. 233–249, 2020.

[32] H. Soh, P. Shu, M. Chen, and D. Hsu, "The transfer of human trust in robot capabilities across tasks," in *Proc. of Robotics: Science and Systems*, 2018.

[33] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *2016 11th ACM/IEEE Intl. Conf. on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 109–116.

[34] W. Zheng, B. Wu, and H. Lin, "Pomdp model learning for human robot collaboration," in *2018 IEEE Conf. on Decision and Control (CDC)*. IEEE, 2018, pp. 1156–1161.

[35] N. Gopalan and S. Tellex, "Modeling and solving human-robot collaborative tasks using pomdps," in *RSS Workshop on Model Learning for Human-Robot Communication*, 2015.

[36] D. Malik, M. Palaniappan, J. F. Fisac, D. Hadfield-Menell, S. Russell, and A. D. Dragan, "An efficient, generalized bellman update for cooperative inverse reinforcement learning," in *35th Intl. Conf. on Machine Learning*, 2018, pp. 3394–3402.

[37] M. Kwon, S. H. Huang, and A. D. Dragan, "Expressing robot incapability," in *2018 ACM/IEEE Intl. Conf. on Human-Robot Interaction*, 2018, pp. 87–95.

[38] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, p. 4–22, 1985.

[39] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[40] F. V. Ferrari, "Cooperative pomdps for human-robot joint activities," Ph.D. dissertation, Normandie, 2017.

[41] T. Hellström and S. Bensch, "Understandable robots-what, why, and how," *Paladyn, Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 110–123, 2018.

[42] S. Sreedharan, T. Chakraborti, C. Muise, and S. Kambhampati, "Expectation-aware planning: A unifying framework for synthesizing and executing self-explaining plans for human-aware planning," in *AAAI*, 2020.

[43] S. Nikolaidis, S. Nath, A. D. Procaccia, and S. Srinivasa, "Game-theoretic modeling of human adaptation in human-robot collaboration," in *2017 ACM/IEEE Intl. Conf. on Human-Robot Interaction*, 2017, pp. 323–331.

[44] Y. Zhang, S. Sreedharan, and S. Kambhampati, "Capability models and their applications in planning," in *2015 Intl. Conf. on Autonomous Agents and Multiagent Systems*, 2015, pp. 1151–1159.

[45] D. S. Syrdal, K. Dautenhahn, K. L. Koay, and M. L. Walters, "The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study," *Adaptive and Emergent Behaviour and Complex Systems*, 2009.

[46] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "Trust Perception Scale-HRI"," in *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016, pp. 191–218.

[47] I. P. Bodala, B. C. Kok, W. Sng, and H. Soh, "Modeling the Interplay of Trust and Attention in HRI: An Autonomous Vehicle Study," in *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 145–147.