A Unifying Framework for Social Motivation in Human-Robot Interaction

Audrey Wang*, Rohan Chitnis*, Michelle Li*, Leslie Pack Kaelbling, Tomás Lozano-Pérez
Massachusetts Institute of Technology
{arwang, ronuchit, limich, lpk, tlp}@mit.edu

Abstract

We develop a framework for *social motivation* in human-robot interaction, where an autonomous agent is rewarded not only by its environment, but also based on the mental state of a human acting within this same environment. We concretize this idea by considering partially observed environments in which the agent's reward function depends on the human's belief. In order to effectively reason about what actions to take in such a setting, the agent must estimate both the environment state and the human's belief. Although instantiations of this idea have been studied previously in various forms, we aim to unify them under a single framework. Our contributions are three-fold: 1) we provide a general POMDP framework for this problem setting and discuss approximations for tractability in practice; 2) we define several reward functions that depend on the human's belief in different ways and situate them with respect to previous literature; and 3) we conduct qualitative and quantitative experiments in a simulated discrete robotic domain to investigate the emergent behavior of, and tradeoffs among, the proposed reward functions.

Introduction

Social motivation influences many of our daily interactions. Broadly, social motivation refers to the idea that people have an intrinsic drive to interact with others and be accepted by them. In human-robot interaction, social motivation can be formulated as a way to incentivize an autonomous agent based on the mental state of a human in the world. Reasoning about another agent's mental state is crucial for practical robotic settings, from self-driving cars to household robots. For instance, we may want to reward a household robot for not only cleaning a kitchen, but also making the human *think* the kitchen is clean, which could require additional communicative actions on the robot's part.

The general idea of *modeling* other agents' mental states has been studied extensively (Wellman 1992; Vogel, Potts, and Jurafsky 2013; Zettlemoyer, Milch, and Kaelbling 2009); however, the notion of *incentivizing* agents based on mental states is relatively understudied (Gray and Breazeal 2012; Talamadupula et al. 2014), and these works typically consider very specific problem settings and incentive structures. For instance, Jaques et al. (2019) study influence-driven incentives, in which agents are rewarded for changing the behavior of other agents through their actions. As another example, Renoux (2015) considers the problem of reasoning for active information-gathering, but does not explore the impact of such objective functions in settings involving multiple actors, such as human-robot interaction. Araya et al. (2010) explicitly describe reward functions that depend on the mental state of an agent, but do not discuss the potential application of these rewards as tools for encouraging social motivation. This work aims to unify all these previous approaches within a single framework that captures social motivation in human-robot interaction.

To concretize the notion of "mental state," we focus on partially observed environments, which require both the agent and the human to select actions based on their current belief (probability distribution over states). We consider

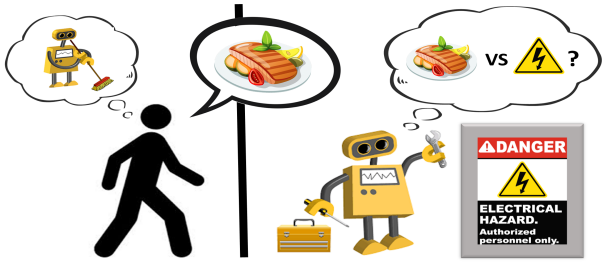

Figure 1: We consider various formulations of *social motivation* in reward function design, by rewarding an agent based on the beliefs of other agents. Suppose we control a household robot working on important electrical repairs. The human, who believes that the robot is simply cleaning the home, asks it to prepare some food. If the robot is equipped with the ability to reason that the human is *not* aware of its current task, and the robot's reward function involves making the human *believe* it is being useful, then it will prioritize its electrical repairs over preparing food and communicate to the human that it is doing something more important. This illustrates how an agent's optimal behavior can be greatly influenced by the impact of the mental states of other agents on its objective. These various behaviors capture different modalities of socially motivated interaction.

* Equal contribution.

settings where the agent's reward function depends on the human's belief in addition to the environment state. Therefore, in order to plan its actions, the agent must estimate not only the environment state, but also the human's belief, requiring the agent to maintain (or approximate) a "belief over beliefs." The reward function's dependence on the human's belief can be realized in various ways, leading to different modes of interesting agent behavior. See Figure 1 for an illustrative example of this idea.

Under this framework, we observe the most interesting behavior when the human and the agent have *heterogeneous* beliefs about the world, which arises from either asymmetric initial knowledge or different observations in an episode.

Our contributions in this work are three-fold:

- We provide a general framework for this problem setting, a partially observable Markov decision process (Kaelbling, Littman, and Cassandra 1998) (POMDP). This framework allows us to directly leverage existing tools for solving POMDPs (Silver and Veness 2010; Somani et al. 2013; Kurniawati, Hsu, and Lee 2008). To make solving these social motivation problems tractable in practice, we discuss approximations in both state estimation (via factoring and discretization) and planning (via online Monte Carlo methods).
- We define several social motivation reward functions that depend on the human's belief in different ways and situate them with respect to literature that propose similar objective functions in related fields, such as single-agent POMDP planning and intrinsic motivation for reinforcement learning.
- We conduct qualitative and quantitative experiments in a simulated discrete robotic room-cleaning domain. In this domain, we investigate the emergent behavior of the different reward functions we propose and discuss their trade-offs both in terms of task-level policy performance and accuracy of the human's belief.

We envision our framework, which focuses on belief recognition, as a precursor to plan and intention recognition. By adding high-level *goals* for agents in the environment, possibly represented as factors in the state, our belief recognition framework can be extended to intent recognition using standard Bayesian inference tools. If an agent is aware of other agents' planning mechanisms, it can simulate planning using approximations of other agents' beliefs, enabling plan recognition. Activity recognition is already present in our current framework's observation model, in which an agent receives observations of another agent's actions.

Related Work

Estimating Mental Models Reasoning about other agents' mental models has been a well-studied problem for decades. The theory of mind framework (Wellman 1992), which describes our ability to reason about the mental states of other people, has roots in psychology and philosophy. It has also been successfully applied to various human-robot interactive settings (Scassellati 2002; Hiatt, Harrison, and Trafton 2011; Devin and Alami 2016). Typically, these approaches estimate the human's belief in order to draw inferences about the plan the human is executing, or the reward function the human is following (Dragan 2017). Some works have addressed the issue of nested (Vogel, Potts, and Jurafsky 2013) or infinitely recursive (Zettlemoyer, Milch, and Kaelbling 2009) belief modeling, which arise in cases where each agent is trying to reason about the other's reasoning process, which in turn depends on its own, etc.

We focus on a two-agent scenario where one of the actors is a human whose actions cannot be controlled. This is in contrast to the field of multi-agent collaboration, notably the decentralized POMDP (Dec-POMDP) (Oliehoek, Cui, and Amato 2016) framework used for decision-making for a team of collaborative agents. The most similar work to ours is the I-POMDP framework (Gmytrasiewicz and Doshi 2005), which seeks to model partially observed environments where the beliefs of both agents are part of the state, as we do. However, our work differs in several key ways: we are considering only a single-agent setting and assume that the human's policy is given, we explicitly study the impact of various reward functions that depend on the human's belief, and our framework is able to handle noise within the agents' perceptions of each other's actions and observations.

Gray and Breazeal (2012) use a self-as-simulator model to estimate the human's belief state. While their work focuses on visual perception, we describe a more general framework where the robot may not necessarily have access to what the human sees. Buehler and Weisswange (2018) consider the problem of online inference of a human's belief state using observable human behavior, but they assume the robot has perfect knowledge about the world; we aim to break down this assumption because our primary point of interest is how an agent can deal with inherent uncertainty over both its own world state estimation *and* the uncertainty of another agent.

Belief-Dependent Rewards The idea of having belief-dependent reward functions within POMDP settings has been studied previously. Araya et al. (2010) proposed an extension to the classic POMDP framework where the agent can be rewarded for having low-entropy beliefs and high expected returns from the environment under their belief (Eck and Soh 2012). It was later shown (Spaan, Veiga, and Lima 2015) that this kind of uncertainty can be modeled directly within the POMDP framework, preserving the theoretical guarantees and existing solvers for POMDPs. Other work has studied rewards that depend on the stability of the belief (Renoux 2015), emotion (Sequeira 2013), influence on other agents' actions (Jaques et al. 2019), and interpretability, i.e. legible planning (Chakraborti et al. 2019).

Social Motivation Framework

In this section, we provide a general POMDP framework for our setting of human-robot interactive social motivation.

Partially observable Markov decision process. An undiscounted POMDP (Kaelbling, Littman, and Cassandra 1998) is a tuple $\langle \mathcal{S}, \mathcal{A}, \Omega, O, T, R, H \rangle$ with: state space $\mathcal{S}$, action space $\mathcal{A}$, observation space Ω, observation model $O(s', a, o) = P(o \mid s', a)$, transition model $T(s, a, s') = P(s' \mid s, a)$, reward function $R(s, a)$, and horizon H. Here, $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, and $o \in \Omega$. The agent selects an action at each timestep, causing the state to transition according to

T and the agent to receive an observation and reward according to O and R. As the state is unobserved by the agent, it must maintain a *belief* $B_A(s)$, a distribution over states (the subscript A stands for "agent"). The objective of the agent is to find a policy π, a mapping from beliefs to actions, that maximizes expected total reward, $\mathbb{E}_\pi\left[\sum_{t=0}^{H} R(s_t, a_t)\right]$, where the expectation is taken over stochasticity in the initial state and transitions. Given an action a and observation o, a Bayes filter provides an exact update expression for the belief: $B'_A(s') \propto O(s', a, o) \sum_{s \in \mathcal{S}} T(s, a, s') B_A(s)$.

Formally, our social motivation model for human-robot interaction is a single-agent POMDP where world states s are pairs $\langle s_E, B_H \rangle$, representing the **E**nvironment state and the belief of the **H**uman. Therefore, the reward function depends on the human's belief, which is unobserved by the agent. This means the agent's belief B_A must estimate a joint distribution over s_E and B_H. The human takes actions in the world and receives observations alongside the agent, which it uses to update its own belief B_H, but does not reason about the agent's state of mind in its decision-making and state estimation. This allows us to circumvent the issue of infinitely nested beliefs (Zettlemoyer, Milch, and Kaelbling 2009).

The complete model is shown in Figure 2. Crucially, the human's belief B_H affects the agent in two ways: directly through the reward function R and indirectly through future states by affecting the actions the human takes.

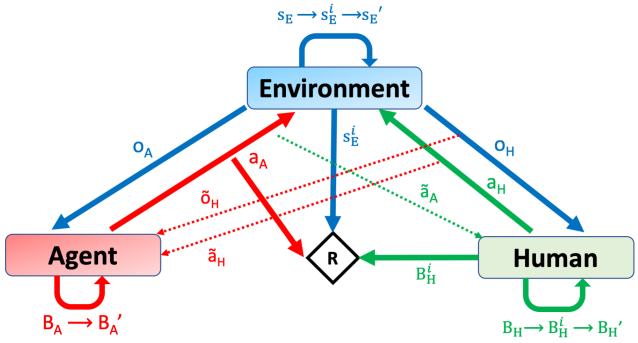

Figure 2: Our model for social motivation in human-robot interaction. From the agent's perspective, the human can be treated as a part of the world, but the agent must reason about the human's belief in order to plan properly, since these beliefs affect the agent through both the rewards R and future states. *Red:* Agent's activities. *Green:* Human's activities. *Blue:* Environment's activities.

In a timestep, the following occur sequentially:

1. The agent selects action $a_A = \pi(B_A)$ using its deterministic policy π. The action a_A is executed, transitioning the state to an intermediate $s_E^i \sim P(s_E^i \mid s_E, a_A)$.
2. The human receives a *corrupted* observation $\tilde{a}_A \sim P(\tilde{a}_A \mid a_A)$ of the agent's action. The agent receives an environment observation $o_A \sim P(o_A \mid s_E^i, a_A)$.
3. The human updates their belief to B_H^i based on $\tilde{a_A}$.
4. The agent receives a reward $R(s_E^i, B_H^i, a_A)$ from the intermediate s_E^i and B_H^i resulting from its action a_A.
5. The human samples an action $a_H \sim \pi_H(B_H^i)$ from their policy π_H. The action a_H is executed in the world, transitioning the state to $s'_E \sim P(s'_E \mid s_E^i, a_H)$.
6. The human receives an environment observation $o_H \sim P(o_H \mid s'_E, a_H)$; the agent receives *corrupted* observations $\tilde{a}_H \sim P(\tilde{a}_H \mid a_H)$ and $\tilde{o}_H \sim P(\tilde{o}_H \mid \tilde{a}_H, o_H)$ of the human's action and observation.
7. The human updates their belief to B'_H based on a_H, o_H.
8. The agent updates its belief to B'_A based on o_A, a_A, $\tilde{o_H}$, and $\tilde{a_H}$ via a (possibly approximate) Bayes filter.

We factor the timestep this way so the agent's reward depends solely on its own action, not on the action of the human. This factorization affords a more intuitive grasp of how the agent's behavior is affected by a belief-dependent reward, since it isolates the effect of the agent's action (see Experiments). Note that we could also generalize the model by defining a timestep as the agent and human acting simultaneously, with transition model $P(s'_E \mid s_E, a_A, a_H)$ and human belief update model $P(B'_H \mid B_H, a_H, o_H, \tilde{a}_A)$. The reward function would then be defined as $R(s'_E, B'_H, a_A)$.

Formally, we simply have a single-agent POMDP setting: from the agent's perspective, the human and their actions can be thought of as part of the environment. Thus, steps 2-7 are just a factored state transition, with the reward function depending on an intermediate state in this transition. The agent's objective is to find a policy π that maximizes expected total reward, $\mathbb{E}_\pi\left[\sum_{t=0}^{H} R(s^i_{E,t}, B^i_{H,t}, a_{A,t})\right]$.

We require the following conditional probabilities: (1) The human's policy $\pi_H(B_H) = P(a_H \mid B_H)$. (2) The human's belief update $P(B'_H \mid B_H, a_H, o_H, \tilde{a}_A)$, defined in this case as two sequential belief updates $P(B^i_H \mid B_H, \tilde{a}_A)$ and $P(B'_H \mid B^i_H, a_H, o_H)$. (3) The environment transition model $P(s'_E \mid s_E, a_A, a_H)$, defined in this case as two sequential transitions $P(s^i_E \mid s_E, a_A)$ and $P(s'_E \mid s^i_E, a_H)$. (4) Agent and human environment observation models $P(o_A \mid s'_E, a_A)$ and $P(o_H \mid s'_E, a_H)$. (5) Corruption models $P(\tilde{a}_A \mid a_A)$, $P(\tilde{a}_H \mid a_H)$, and $P(\tilde{o}_H \mid \tilde{a}_H, o_H)$.

We note that this model can flexibly handle a variety of assumptions that could be made for tractability or practical applicability. For instance, the term $P(B'_H \mid B_H, a_H, o_H, \tilde{a}_A)$ in principle captures a distribution over all possible human beliefs; in practice, it can be easier to assume that the human uses a Bayes filter: $P(B'_H \mid B_H, a_H, o_H, \tilde{a}_A) = \mathbb{1}[B_H \xrightarrow{a_H, o_H, \tilde{a}_A} B'_H]$. Also, if one is using an environment where it is not sensible for the agent and human to see each others' actions and observations, then the corruption models can be set to uniform, thus carrying no information.

In this work, we are not considering the model-learning problem (Shani, Brafman, and Shimony 2005; Doshi and Roy 2007); we assume all the conditional probabilities are given and instead focus on 1) approximations for tractable planning and 2) understanding the emergent behavior under different instantiations of the reward function.

Tractability. We briefly discuss approximations for state estimation and planning to make this framework practical.

We leverage two key ideas for approximate state estimation: factoring and discretization. Instead of maintaining a joint belief B_A over both s_E and B_H, we estimate B_A as a product of marginals over s_E and B_H, tracked separately. Furthermore, to represent the belief over B_H, which in theory is a distribution over all possible human beliefs, we discretize the probability space over each factor in the state, and maintain B_A only over this coarser granularity. Of course, this causes loss in precision, and may cause arbitrarily bad performance in adversarially designed domains. This discretization approach can be made more robust by employing a tile coding scheme (Sherstov and Stone 2005), which we leave for future work. We discretize further by employing particle filters (Djuric et al. 2003) to estimate s_E and B_H.

For planning, we employ standard online POMDP solution strategies (Silver and Veness 2010; Somani et al. 2013; Bonet and Geffner 2000), which estimate the value function from Monte Carlo rollouts to give the estimated best next action to take; the planner is run on every timestep.

Reward Functions for Social Motivation

In this section, we describe several ways to instantiate the reward function $R(s_E^i, B_H^i, a_A)$ from the framework detailed in the previous section, drawing on the notion of social motivation. In doing so, we aim to unify existing literature in related fields and showcase the generality of our framework.

To begin, suppose that the environment defines a reward function, which we call the *task-level* reward: $\bar{R}(s_E^i, a_A)$.

Task-Only. The simplest instantiation of a reward function in our framework just ignores the human's belief:

$$R(s_E^i, B_H^i, a_A) = \bar{R}(s_E^i, a_A). \qquad \textit{(Task-Only)}$$

Therefore, the agent is incentivized only to perform as well as possible at its task, and only cares about the human's belief to the extent that the human's actions affect the environment state. In many settings, it can be useful to reward the agent explicitly based on properties of the human's belief.

Human-Expectation. Drawing on this intuition, we next consider a reward that asks the agent to not only perform well at its task, but also make the human *believe* it is doing so (under the same task-level reward function $\bar{R}$):

$$R(s_E^i, B_H^i, a_A) = \bar{R}(s_E^i, a_A) + \lambda \mathbb{E}_{\bar{s}_E^i \sim B_H^i}\left[\bar{R}(\bar{s}_E^i, a_A)\right]. \qquad \textit{(Human-Expectation)}$$

Here, the parameter $\lambda > 0$ controls the relative importance of these two objectives. This objective will typically lead to *explanatory* behavior, in which the agent describes why it is choosing to perform certain actions, so that the human understands the context behind the agent's decisions. It is useful in settings such as household robotics, where a robot helper should make sure the human believes it is performing its tasks properly. Note that in this reward formulation, we are assuming that the agent's action a_A is perfectly seen by the human. This optimistic assumption could easily be replaced by slight variants, such as taking an expectation under the corruption model $P(\tilde{a}_A \mid a_A)$.

This *Human-Expectation* reward is adapted from a reward proposed by Araya et al. (2010) for planning with rewards that depend on an agent's own belief.

Human-Certainty. Next, we have a reward function that asks the agent to perform well at its task while reducing uncertainty in the human's belief:

$$\begin{aligned} R(s_E^i, B_H^i, a_A) &= \bar{R}(s_E^i, a_A) - \lambda S(B_H^i) \\ &= \bar{R}(s_E^i, a_A) + \lambda \mathbb{E}_{\bar{s}_E^i \sim B_H^i}\left[\log B_H^i(\bar{s}_E^i)\right]. \end{aligned} \qquad \textit{(Human-Certainty)}$$

Here, the notation $S(B_H^i)$ refers to the Shannon entropy of the human's belief, and is highest when this belief is diffuse, leading to lower rewards. Again, $\lambda > 0$ trades off between these two terms in the objective. This objective will typically lead to communication of *declarative* information, where the agent transmits facts about the state in order to reduce the human's uncertainty about it (assuming one has taken care to design the action space such that transmitting incorrect information is disallowed). This reward function is useful in settings such as driving; a human driver will often start slowing down early on at a traffic jam to implicitly communicate the situation ahead to the driver behind them.

This *Human-Certainty* reward is also adapted from a reward proposed by Araya et al. (2010), and can be easily combined with the previous one. Note that this reward function can be adapted to adversarial settings by choosing $\lambda < 0$, thus incentivizing obfuscation of the human's belief.

The remaining beliefs we study depend on both B_H and B_H^i: the human's belief before and after the robot acts, respectively. This requires a small change to the framework described in the previous section; we will not burden notation here by redefining the relevant terms.

Influence. We next consider a reward that encourages the agent to take actions resulting in a large change in the human's policy, measured via the KL-divergence:

$$\begin{aligned} &R(s_E^i, B_H, B_H^i, a_A) \\ &\quad = \bar{R}(s_E^i, a_A) + \lambda D_{\mathrm{KL}}(\pi_H(B_H) \parallel \pi_H(B_H^i)) \\ &\quad = \bar{R}(s_E^i, a_A) + \lambda D_{\mathrm{KL}}(P(a_H \mid B_H) \parallel P(a_H \mid B_H^i)). \end{aligned} \qquad \textit{(Influence)}$$

Here, $\lambda > 0$ is a tradeoff parameter. Recall that the human updates their belief on the basis of their perception of the agent's action, and also indirectly through the influence the agent's action has on the environment state. Therefore, this objective incentivizes the agent to act in ways that cause the human's belief update to alter their policy. This reward function is useful, for instance, in situations where a robot must strive to get the attention of a human operator or teammate, such as to warn them about something in the environment that they do not know about, to get them to change their behavior. Note that one could also set $\lambda < 0$ to incentivize the agent to keep the human's policy as stable as possible.

A similar reward function, which serves as the inspiration for this *Influence* reward, was studied in the context of intrinsic motivation for deep reinforcement learning by Jaques et al. (2019). Their experiments show that motivating agents in a multi-agent system to alter each others' policies as much as possible provides a useful exploratory bias for coordination games, leading to more sample-efficient learning.

Human-Stability. Finally, we consider a reward that encourages the agent to keep the human's belief stable:

$$R(s_E^i, B_H, B_H^i, a_A) = \bar{R}(s_E^i, a_A) - \lambda D_{\text{KL}}(B_H \parallel B_H^i). \quad \textit{(Human-Stability)}$$

Again, $\lambda > 0$ is a tradeoff parameter. This objective incentivizes the agent to act in ways that cause low-magnitude updates to the human's belief, and thus will typically lead to the agent taking less surprising or risky behavior when completing a task. This can be useful in factory settings where humans are often nearby robots performing tasks; these robots should take care to act in ways that do not cause the humans to suddenly believe that they are broken or behaving poorly.

This *Human-Stability* reward function is inspired by a similar formulation that was studied by Renoux (2015) within an information-gathering setting. In that work, the KL-divergence between an agent's old and new beliefs is used as a measure of novelty of an observation; if an observation is predictable under the belief, it will not affect the belief too much, and thus has lower expected novelty. Here, we instead study the utility of this reward formulation as a tool for social motivation.

Note that all of the reward functions described in this section are the true reward functions of the unobserved POMDP state. In practice, the agent will optimize an expectation over these rewards with respect to its current belief.

Experiments

We aim to understand the behavior induced by the various reward functions, via qualitative and quantitative experiments in a simulated discrete robotic room-cleaning domain. In this domain, we investigate the emergent behavior of the reward functions we propose, and discuss their tradeoffs both in terms of task-level policy performance and accuracy of the human's belief. We find that the reward functions produce a suite of interesting behavior to analyze, and that our approximation strategies allow our framework to scale reasonably well to larger domains.

Our experiments are designed to answer the following:

- What kind of emergent behavior results from each of our proposed social motivation reward functions in practice?
- How does each reward function impact 1) the typical accuracy of the true human belief and 2) the task-level rewards received by the agent?
- How well does our framework scale to larger domains?

Domain Description

We conduct our experiments on a discrete household robot domain where the environment is a gridworld representing a house with n bedroom locations: L_1, L_2, …, L_n. The state space is as follows: let each $s \in \mathcal{S}$ consist of n state variables $\{L_1, L_2, ..., L_n\}$, which denote the binary-valued *cleanliness* (either clean or dirty) of the rooms. See Figure 6 (left) for a visualization of this domain when $n = 15$.

The agent is a household cleaning robot whose set of actions is $\mathcal{A}_A = \{c_1, c_2, ..., c_n\}$; at each timestep, the robot can choose an action $a_A \in \mathcal{A}_A$ to clean one of the rooms in the house (action c_i cleans bedroom L_i). After its action, the robot receives a (noisy) observation o_A about whether it successfully cleaned the room, and the human receives a (noisy) observation $\tilde{a}_A$ about the robot's action.

The human chooses an action from the set $\mathcal{A}_H = \{d_1, d_2, ..., d_n\}$, where each action $a_H \in \mathcal{A}_H$ denotes *dirtying* one of the rooms in the house. After each action, the human receives a (noisy) observation o_H about whether it dirtied the room, and the robot receives (noisy) observations $\tilde{a}_H$ and $\tilde{o}_H$ about the human's action and observation.

Within a timestep, the robot's action c_i can stochastically transition room L_i from dirty to clean, and the human's action d_j can stochastically transition room L_j from clean to dirty. Additionally, a clean room that is not acted on by the human or robot has some probability of becoming dirty.

The robot receives positive task-level rewards for cleaning dirty rooms (greater for higher-numbered rooms) and negative task-level rewards for cleaning already-clean rooms.

Experimental Setup

Human Policy and Belief Update The human's policy π_H is deterministic: the human will choose to act upon the room that has the highest probability of being clean, under their belief B_H. In case of a tie, the human will choose to act upon the highest-numbered room.

We model the human belief update as a Bayes filter. Since π_H is deterministic, we have that $a_H = \pi_H(B_H)$.

Approximation Methods As the states of the rooms are independent, the human and robot beliefs about each state s are a set of n probabilities where entry i corresponds to the probability that room L_i is clean.

The robot's belief over the human's belief is approximated as a set of particles; we update this distribution via particle filtering (Djuric et al. 2003). Furthermore, for the robot's belief we discretize the space of possible beliefs the human could have about each room (a number between 0 and 1) into ten buckets, and maintain only a discrete distribution over these ten buckets. We use POMCP (Silver and Veness 2010), an online POMDP planning algorithm, to choose actions. In this process, the robot has only sample access to the models and reward function, rather than the complete distributions.

Model Descriptions In the environment transition model, the robot's action a_A successfully cleans a dirty room with probability 0.9, while the human's action a_H dirties a clean room with probability 0.5. In addition, any clean room that is not acted on by either becomes dirty with probability 0.2. In the environment observation model, the robot (resp. human) receives the correct observation about the state of the room it is in with probability 0.8 (resp. 0.9). The corruption models are as follows. The robot receives correct observations $\tilde{a}_H$ and $\tilde{o}_H$ about the human's action and observation with probability 0.85 for each; the remainder is split uniformly among all other possibilities. The human receives the correct observation $\tilde{a}_A$ about the robot's action with probability 0.9, again with the remainder split uniformly.

Qualitative Results and Discussion

In order to clearly understand the emergent behavior resulting from each reward function, we perform qualitative

Figure 3: Qualitative results of emergent behavior for each reward function. See text for detailed descriptions.

analysis in a simple environment with $n = 3$ rooms. All experiments use the same initial state: room 2 is clean, while rooms 1 and 3 are dirty; the robot initially incorrectly believes that all the rooms are dirty; the human initially believes that room 2 is clean with probability 0.7 and rooms 1 and 3 are almost surely dirty. Crucially, in these experiments, the human and robot initially start out with heterogeneous beliefs, and it is this asymmetry in their beliefs that allows us to observe interesting emergent behavior in the early timesteps of episodes in these experiments.

Each episode consists of a sequence of 10 timesteps executed from this initial state. In Figure 3, we display the first four timesteps for each reward function. Each timestep shows the state after the robot and the human have acted.

Using the *Task-Only* reward function (a), the robot cleans the dirtiest room (3) on timestep 1. On the same timestep, the human enters the room and makes Room 3 dirty again, but receives the observation that it is actually clean. The robot observes the human's observation and now believes Room 1 and Room 2 are dirty, and that Room 3 probably is clean. It thus chooses to clean Room 2 on timestep 2, *ignoring that the human believes Room 2 is already clean*.

With the *Human-Expectation* reward (b), the robot diverges on timestep 2, now choosing to clean Room 1 instead of 2, since *under the human's belief*, Room 2 is clean. On timestep 3, the robot and human both believe that Room 3 has the highest probability of being dirty. Thus, the robot chooses to clean Room 3.

The *Human-Certainty* reward function (c) encourages the robot maximize the human's certainty about the world state. Since the human starts off certain that Room 1 and 3 are dirty, but only believes with 0.7 probability that Room 2 is clean, the robot chooses to clean Room 2, and in doing so provides the human with more information about the state of that room. Due to the stochasticity in transitions and observations, the human's belief will never collapse to complete certainty, so the robot repeatedly cleans Room 2 to increase the human's certainty about Room 2's state.

Under the *Influence* reward function (d), the robot attempts to take actions that alter the human's behavior. The human initially believes that Room 2 is clean, and thus would typically move to Room 2 next, because their policy is to move to the room they believe has the highest likelihood of being clean. In order to change this behavior, the robot cleans Room 3. Similar to (a) and (b), the human observes (incorrectly) that Room 3 is clean, and thus would typically choose to stay in Room 3 on timestep 2. Thus, the robot once again decides to alter the human's behavior by cleaning Room 1 in timestep 2. This "cat-and-mouse" behavior is a direct consequence of the *Influence* reward.

Finally, the *Human-Stability* reward function (e) encourages the robot to keep the human's belief as stable as possible. Since the human already believes with 0.7 probability that Room 2 is clean, the robot chooses to clean Room 2 again to cause the least amount of change in the human's belief. Similar logic follows for the remaining timesteps.

In summary, the various formulations of belief-dependent reward functions result in the expected emergent behavior for this $n = 3$ domain setting. We are able to unify works that aim to produce specific types of socially motivated behavior; the qualitative results presented suggest that all these behaviors can be realized by changing only the reward function, within a single unified framework.

Quantitative Results and Discussion

We conduct quantitative experiments to investigate the effect of each reward function on both task-level reward and human belief accuracy at various settings of λ, the reward function trade-off parameter. We use the same $n = 3$ environment and initial state as in the qualitative experiments. Each episode consists of 3 timesteps.

Figure 4 shows the effect of each reward function on total task-level reward for the 3 timesteps, while Figure 5 illustrates the effect on the average accuracy of the true human belief. We use the basic task-only reward function (blue curve) as a control. There are several interesting trends here:

1) For *Human-Certainty* and *Human-Stability*, λ has a negative effect on task-level reward but a positive effect on the true human belief's accuracy. Intuitively, if certainty and stability are weighted more heavily, the robot is incentivized to be redundant in the interest of clarity. This redundancy by the robot leads to higher accuracy of the human's belief, at the expense of task-level rewards.

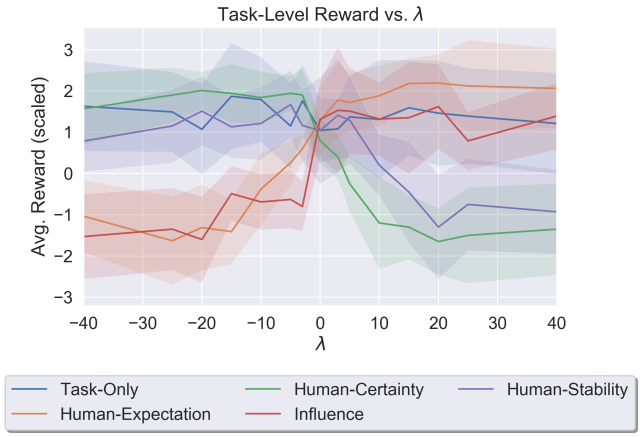

Figure 4: Effect of λ (trade-off parameter) on task-level (environment) reward for each reward function. Weighting the certainty or stability of the human's belief too highly lowers task-level rewards (green & purple vs. blue), while acting in a way that the human expects is beneficial improves task-level rewards (orange vs. blue).

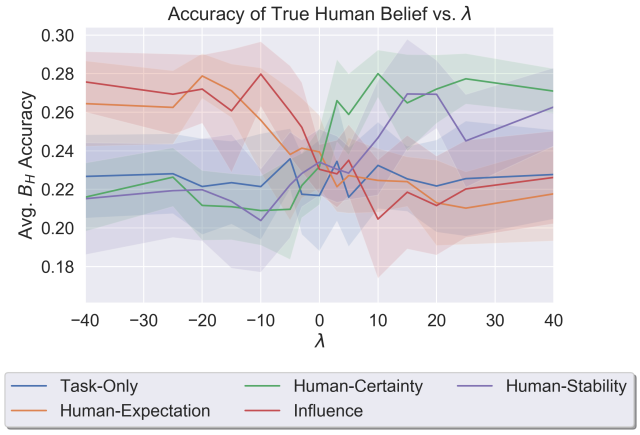

Figure 5: Effect of λ (trade-off parameter) on the accuracy of the true human belief (relative to the true environment state) for each reward function. The human's belief tends to be most accurate when the robot is incentivized to make it have low entropy, like in the certainty and stability rewards (green & purple vs. rest).

2) For the *Human-Expectation* and *Influence* rewards, λ has a positive effect on task-level reward. In the initial state, the human's belief is more accurate than the robot's, and so the robot receives higher task-level reward by acting in ways that the human believes to be beneficial. Similarly for *Influence*, a positive λ discourages the robot from taking the same action repeatedly. This has a positive impact on task-level reward because the robot is less likely to incur the penalty for cleaning an already-clean room.
3) For the *Influence* reward, λ has a negative effect on the accuracy of the human's belief. This is because if the robot tries to influence the human's actions, it needs to change the human's belief sharply, which generally leads to the human having higher uncertainty about the state.
4) The task-level reward achieved with the *Human-Expectation* reward function is higher than that of the baseline *Task-Only* reward function, for positive values of λ. This is because the human's belief is initially more accurate than the robot's. Even in this simple environment, we can see that considering both agents' beliefs is more robust to the error in any single agent's individual belief.

Finally, we show that our implementation, even with its "belief over beliefs" estimation, scales reasonably with increasing domain size (Figure 6). While the performance can still be improved, our framework demonstrates tractability without incurring huge losses in accuracy of information.

Limitations and Future Work

A major limitation of the current approach is that in order to allow our method to scale up to medium-sized domains, we found it necessary for the agent's belief over the human's belief to reason over a coarse discretization of the human's belief space. Of course, such a method does not scale well to higher-dimensional beliefs, but was sufficient for the factored domain we considered in our experiments.

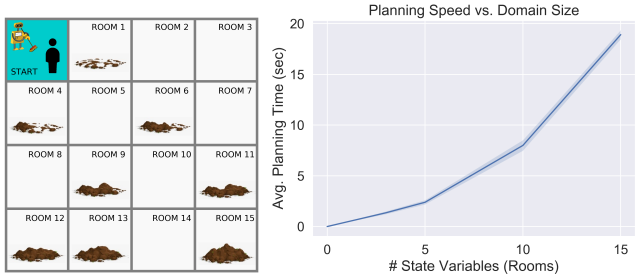

Figure 6: *Left:* Visualization of the domain with $n = 15$ rooms. The robot receives higher reward for cleaning rooms that have more dirt in them. *Right:* Average planning time (per timestep) versus the size of the domain. We can see that planning time scales reasonably with increasing domain size, up to medium-sized domains.

Future work should investigate better representations of the agent's belief, via distributions whose support affords better coverage of the underlying space of human beliefs. For instance, one possibility is to treat the human's belief as parameterized by a set of latent variables, which could represent parameters of a known distribution or weights of a neural network, and keeping a distribution over these latents.

Another avenue for future work is to use other approximate POMDP planning strategies such as DESPOT (Somani et al. 2013) to further improve scalability. We would expect to still be able to observe interesting behavior emerging from our social motivation rewards in more complex domains.

It would also be interesting to explore the notion of reasoning over possible *goals* that the human may have, rather than only beliefs as we do here. In principle, the agent could have uncertainty about beliefs, desires, and intentions; reasoning about all of these together is a necessary step toward enabling practically useful decision-making. This reasoning would also act as an avenue for practical plan and intention recognition in human-robot collaborative settings.

References

Araya, M.; Buffet, O.; Thomas, V.; and Charpillet, F. 2010. A pomdp extension with belief-dependent rewards. In *Advances in neural information processing systems*, 64–72.

Bonet, B., and Geffner, H. 2000. Planning with incomplete information as heuristic search in belief space. In *Proceedings of the Fifth International Conference on Artificial Intelligence Planning Systems*, 52–61.

Buehler, M., and Weisswange, T. 2018. Online inference of human belief for cooperative robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 409–415.

Chakraborti, T.; Kulkarni, A.; Sreedharan, S.; Smith, D. E.; and Kambhampati, S. 2019. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. *Twenty-Ninth International Conference on Automated Planning and Scheduling (ICAPS 2019)*.

Devin, S., and Alami, R. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*, 319–326. IEEE.

Djuric, P. M.; Kotecha, J. H.; Zhang, J.; Huang, Y.; Ghirmai, T.; Bugallo, M. F.; and Miguez, J. 2003. Particle filtering. *IEEE signal processing magazine* 20(5):19–38.

Doshi, F., and Roy, N. 2007. Efficient model learning for dialog management. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, 65–72. ACM.

Dragan, A. D. 2017. Robot planning with mathematical models of human state and action. *CoRR* abs/1705.04226.

Eck, A., and Soh, L.-K. 2012. Evaluating pomdp rewards for active perception. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, 1221–1222. International Foundation for Autonomous Agents and Multiagent Systems.

Gmytrasiewicz, P. J., and Doshi, P. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research* 24:49–79.

Gray, J., and Breazeal, C. 2012. Manipulating mental states through physical action. In Ge, S. S.; Khatib, O.; Cabibihan, J.-J.; Simmons, R.; and Williams, M.-A., eds., *Social Robotics*, 1–14. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hiatt, L. M.; Harrison, A. M.; and Trafton, J. G. 2011. Accommodating human variability in human-robot teams through theory of mind. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 3040–3049.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101:99–134.

Kurniawati, H.; Hsu, D.; and Lee, W. S. 2008. SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, volume 2008. Zurich, Switzerland.

Oliehoek, F. A.; Cui, B.; and Amato, C. 2016. *A concise introduction to decentralized POMDPs*. Springer.

Renoux, J. 2015. *Contribution to multiagent planning for active information gathering*. Ph.D. Dissertation, Örebro University.

Scassellati, B. 2002. Theory of mind for a humanoid robot. *Autonomous Robots* 12(1):13–24.

Sequeira, P. 2013. Socio-emotional reward design for intrinsically motivated learning agents. *Unpublished doctoral dissertation). Universidada Técnica de Lisboa*.

Shani, G.; Brafman, R. I.; and Shimony, S. E. 2005. Model-based online learning of pomdps. In *European Conference on Machine Learning*, 353–364. Springer.

Sherstov, A. A., and Stone, P. 2005. Function approximation via tile coding: Automating parameter choice. In *International Symposium on Abstraction, Reformulation, and Approximation*, 194–205. Springer.

Silver, D., and Veness, J. 2010. Monte-carlo planning in large POMDPs. In *Advances in neural information processing systems*, 2164–2172.

Somani, A.; Ye, N.; Hsu, D.; and Lee, W. S. 2013. DESPOT: Online POMDP planning with regularization. In *Advances in neural information processing systems*, 1772–1780.

Spaan, M. T.; Veiga, T. S.; and Lima, P. U. 2015. Decision-theoretic planning under uncertainty with information rewards for active cooperative perception. *Autonomous Agents and Multi-Agent Systems* 29(6):1157–1185.

Talamadupula, K.; Briggs, G.; Chakraborti, T.; Scheutz, M.; and Kambhampati, S. 2014. Coordination in human-robot teams using mental modeling and plan recognition. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2957–2962. IEEE.

Vogel, A.; Potts, C.; and Jurafsky, D. 2013. Implicatures and nested beliefs in approximate decentralized-pomdps. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 74–80.

Wellman, H. M. 1992. *The child's theory of mind.* The MIT Press.

Zettlemoyer, L.; Milch, B.; and Kaelbling, L. P. 2009. Multi-agent filtering with infinitely nested beliefs. In *Advances in neural information processing systems*, 1905–1912.