Social Interactions as Recursive MDPs

Ravi Tejwani[1*], Yen-Ling Kuo[1*], Tianmin Shu[1],Boris Katz[1],Andrei Barbu[1]

***Abstract*— While machines and robots must interact with humans, providing them with social skills has been a largely overlooked topic. This is mostly a consequence of the fact that tasks such as navigation, command following, and even game playing are well-defined, while social reasoning still mostly remains a pre-theoretic problem. We demonstrate how social interactions can be effectively incorporated into MDPs by reasoning recursively about the goals of other agents. In essence, our method extends the reward function to include a combination of physical (something agents want to accomplish in the configuration space, a traditional MDP) and social intentions (something agents want to accomplish relative to the goals of other agents). Our S-MDPs (social MDPs) allow specifying reward functions in terms of the estimated reward functions of other agents, modeling interactions such as helping or hindering another agent (by maximizing or minimizing the other agent's reward) while balancing this with the actual physical goals of each agent. Our formulation allows for an arbitrary function of another agent's estimated reward structure and physical goals, enabling more complex behaviors such as politely hindering another agent or aggressively helping them. Extending S-MDPs in the same manner as I-POMDPs extension would enable interactions such as convincing another agent that something is true. To what extent the S-MDPs presented here and their potential S-POMDPs variant account for all possible social interactions is unknown, but having a precise mathematical model to guide questions about social interactions both has practical value (we demonstrate how to make zero-shot social inferences and one could imagine chatbots and robots guided by S-MDPs) and theoretical value by bringing the tools of MDP that have so successfully organized research around navigation to hopefully shed light on what social interactions really are given their extreme importance to human well-being and human civilization.**

I. INTRODUCTION

Progress on modeling social interactions and giving machines social goals, such as being particularly nice to a user, is significantly hampered by the lack of theoretical models which characterize what social interactions are. Great practical progress was made in robot navigation and later in sensing with the introduction of MDPs [1] and POMDPs [2]. Defining the problem clearly allowed us as a field to understand what we can model and how to do so. Until we take this same step for social interactions they will remain on shaky ground and despite their importance to virtually every interaction humans engage in they will remain largely off-limits to machines.

We introduce an extension of MDPs, which we term Social MDPs or S-MDPs. To do so, we make several assumptions. First, that agents have physical goals and social intentions, and their overall reward structure is some arbitrary combination of the two, potentially accompanied by other terms. Physical goals are what MDPs can already express, some of function of points in a configuration configuration space. Social intentions are a function of the estimate of the reward structure of another agent. For example, a reward that hinders another agent is some negative function of the estimated reward of that agent. Complicating matters is the fact that social rewards like beliefs can be recursive, an agent may want to help another agent help them. To model this, S-MDPs are recursive up to some depth, much like interactive POMDPs [3] (I-POMDPs). Unlike I-POMDPs, S-MDPs are not recursive in terms of agent's beliefs about the state of the world. Instead, S-MDPs are recursive in terms of the rewards of the agents. This makes S-MDPs and I-POMDPs orthogonal and complementary. S-MDPs are specifically formulated to not interfere with the standard extension from MDPs to POMDPs, making partial observability trivial to include. While we do not develop a joint SI-POMDP here, this is a straightforward extension which would cover far more of the space of social interactions, although one that is computationally challenging.

Our contributions are:

1) Formulating Social MDPs where an agent's reward function is an arbitrary function of the recursive estimate of another agent's reward and a physical goal.
2) An implementation where that function is a linear transformation, which captures notions of helping and hindering
3) Experimental validation of zero-shot social understanding where agents that have never been asked to help or hinder do so.

The space of social interactions which can be captured by S-MDPs is unknown, largely because the space of possible social interactions is ill-defined at present with many proposed mutually-incompatible and incomplete taxonomies. In the future, we intend to validate S-MDPs with human subjects experiments to characterize which interactions are representable as S-MDPs, S-POMDPs, and SI-POMDPs.

II. RELATED WORK

A. Inferring Social Interactions

Social interactions in multi-agent setting has been explored in previous work in the form of estimating social goal of the agent using theory based models for goal attribution [4–8], Bayesian inverse planning to infer agent's goal given the observations of their behavior [9, 10] and co-ordination for human-AI collaboration [11, 10, 12, 13].

Social exchange, built upon the Piaget's Theory of Social Exchanges [14], was proposed as a value for performing a

∗ Equal contribution
[1]Computer Science and AI Laboratory, MIT
{tejwanir,ylkuo,tshu,boris,abarbu,}@mit.edu

service or the satisfaction value for receiving it. It used hybrid BDI-POMDP agent models defined over the BDI (Beliefs, Desires, Intentions) with plans derived from POMDP using social exchange strategies [15, 16].

Social interactions has also been explored in videos of the group activities where people engage in social activities such as walking, waving, hugging, hand shaking [17–19]. These videos demonstrate the social relationship between people (agents) and the snippets are further used to predict the overall relationship. In contrast, we focus on recursively modelling the agents to estimate each others' social intentions at different levels of reasoning.

B. Simulating Agent Trajectories for Social Interactions

The simulation of agent trajectories for social interaction task was first demonstrated by [20] through the set of animations involving the movements of geometrical figures. They conducted human experiments to investigate their perception of the social interaction between the geometrical figures. Simulating agent behaviors in physics engine had been conducted by collecting and validating the datasets on social perception tasks in fully observable environments [21, 22] and in partially observable environments [23, 11].

These frameworks focused on (a) each agent having either a physical goal or a social goal, and (b) on the entire social relationship between the two agents. However, in our framework we consider each agent to have their physical goal as well as a social intention towards the other agent rather than the social relationship as a whole. Both the goal and social intention are recursively estimated by each agent at each time step as they take their corresponding actions towards the goal.

C. Interactive POMDPs

Interactive POMDP frameworks [24–26] were proposed as an extension to POMDP in which an agent attempts to model the other agent in incorporating beliefs about the other agent in terms of preferences, capabilities, and beliefs into nested levels (interactive beliefs). The interactive beliefs were maintained over interactive states which included the physical states and the models of other agents behaviors. The recursive Bayesian update was used to maintain the beliefs over time such that the solution maps the belief states to actions. Furthermore, [27] and [24] described the approximate solutions to I-POMDP using Interactive Particle Filtering for descending the levels of the interactive belief hierarchies and samples that propagates the interactive beliefs at each level.

S-MDP takes similar recursive inference idea as I-POMDP. But rather than inferring actions through beliefs, we estimate the other agents' goal and social intentions at each level and use the estimated goal and social intentions to infer the policy of the other agent at each level.

III. Social MDP

This section formulates the Social MDP (S-MDP) as a two-player Markov game inspired by cognitive hierarchy models of games [28] and nested MDP models [29–31]. The S-MDP framework consists of iterative decision making for each agent doing *l*-level reasoning of other agents' social intentions. Each agent plans its optimal policy by assuming that the other agent's policy is based on lower levels of reasoning. This results in a finitely nested MDP where at each level the agent needs to choose policy that maximizes its own reward with respect the policy of the other agent conditioned on the estimated social intention. We refer this estimated policy of the other agent as *social intention policy* in this paper.

A. Assumptions

As in the typical MDP setting, the states are fully observable to both agents. Both agents have full access to the underlying MDP except for the social intention of the other agent. Each agent has to estimate the social intention of the other agent while planning its own action at each time step. While we present the S-MDP with two agents, this framework can be extended to any number of agents.

B. S-MDP Formulation

We consider a multi-agent system in which one agent determines its optimal policy by considering the policy of the other agent at l levels of reasoning. At $l > 0$, the agent computes its optimal policy based on the other agent's policy at lower levels $0, 1, \cdot, l-1$. At $l = 0$, it solves a typical MDP. A S-MDP for an agent i at level l is defined as:

$$M_i^l = \langle \mathcal{S}, \mathcal{A}, T, \chi_{ij}, R_i, \gamma \rangle \tag{1}$$

where

- S is a set of states in the environment where $s \in \mathcal{S}$.
- $\mathcal{A} = \mathcal{A}_i \times \mathcal{A}_j$ is the set of joint moves of all agents. $a_i \in \mathcal{A}_i$ and $a_j \in \mathcal{A}_j$ are the actions for agent i and j respectively.
- T denotes the probability distribution of going from state $s \in \mathcal{S}$ to next state $s' \in \mathcal{S}$ given actions of all agents: $T\left(s' \mid s, a_i, a_j\right)$.
- χ_{ij} represents the social intention of agent i towards agent j and is used in reward function to define the reward in helping/hindering the other agent j.
- R_i is the reward function for agent i that maps the state, joint actions, and its social intentions towards the other agent to real numbers.
- γ is a discount factor: $\gamma \in (0, 1)$.

a) Reward: Each agent can have its own physical goal, e.g. going to a landmark, as well as the social goals, i.e. helping or hindering other agents. The immediate reward of a social agent i is characterized by its social intention towards the other agent j as follows:

$$R_i(s, a_i, a_j, \chi_{ij}) = r\left(s, a_i, g_i\right) + \chi_{ij} \cdot r(s, a_j, g_j) - c(a_i) \tag{2}$$

where $r(\cdot)$ is the static reward given the agent's own physical goal (e.g. g_i and g_j); $c(\cdot)$ is the cost for taking an action; χ_{ij} indicates the social intention towards agent j showing how much agent i would like to help/hinder agent j. When $\chi_{ij} > 0$, agent i tends to help; when $\chi_{ij} < 0$, agent i tends

to hinder; and $\chi_{ij} = 0$ means agent i is neutral to the other agent j. In this setting, a social agent can maximize its reward if it successfully helps or hinder other agents.

b) Estimating goals and social intentions: To solve an agent's MDPs over different levels, it needs to estimate the other agent's physical goal and social intentions at different levels of reasoning. Similar to [21], the physical goal g_j of agent j is predicted by i using the Bayes's rule:

$$P(g_j|s^{1:T}) \propto \int_{\widetilde{\chi}_{ji}} P(s^{1:T}|g_j, \widetilde{\chi}_{ji}) \cdot P(g_j) \cdot P(\chi_{ji})\ d\widetilde{\chi}_{ji} \quad (3)$$

Since the agent is estimating the social intention at the same time, the estimation of physical goal needs to marginalize over the estimated social intention as well. The social intention of agent i towards agent j estimated by agent k at level l is denoted as $\tilde{\chi}_{ij}^{k,l}$. In the two-player setting, k can be either agent i or j depending on which agent is making estimation. We will describe how to update the estimate of social intention in Section III-D. When solving agent i's MDP at level l, this estimated social intention is further used to compute the other agent j's social intention policy $\widetilde{\psi}_j^{i,l} : \mathcal{S} \times \mathcal{A}_j \times \widetilde{\chi}_{ji}^i \to [0, 1]$, i.e. $P(a_j \mid s, \chi_j)$.

C. Planning for S-MDP

Analogous to MDP, the state-action value is the sum of immediate reward and the expected value in the future. Since the agent i is interacting with agent j, it needs to estimate what actions agent j may take to compute its state-action value. S-MDP considers the expectation over the estimated social intention of agent j in the Q function:

$$Q_i^l(s, a_i, a_j, \chi_{ij}) = R(s, a_i, \chi_{ij}) + \gamma \sum_{s' \in S} T\left(s, a_i, a_j, s'\right) \sum_{a_i'} \sum_{a_j'} \int_{\widetilde{\chi}_{ji}^i} Pr(\tilde{\chi}_{ji}^i \mid s, a_i) \widetilde{\psi}_j^{i,l}(s', a_j', a_i', \widetilde{\chi}_{ji}^i) Q_i^l(s', a_i', a_j', \chi_{ij})\ d\widetilde{\chi}_{ji}^i \quad (4)$$

The l-level social intention policy $\widetilde{\psi}_j^{i,l}$ of the agent j is predicted by i using the Q function at level l-1:

$$\widetilde{\psi}_j^{i,l}(s, a_j, a_i, \chi_{ji}) = \frac{\exp(Q_j^{l-1}(s, a_i, a_j, \chi_{ji})/\tau)}{\sum_{a_i} \sum_{a_j} \exp(Q_j^{l-1}(s, a_i, a_j, \chi_{ji})/\tau)} \quad (5)$$

This is a softmax policy where τ is the temperature parameter controlling how much the agent j follows the greedy actions. Based on Eq. 4, in order to use agent j's Q function at level l-1, it requires to compute agent i's Q function at level l-2, and so on. This involves solving recursive MDPs at levels $0, 1, \cdots, l$-1.

D. Social Intention Update

[10] showed that humans can easily estimate agents' social intention by watching agents carrying out their actions. The confidence of such estimation increases as they observe actions for more time steps. Taking this idea, in S-MDP, an agent's estimation of the other agent's social intention at time step t is updated based on the actions performed by the agents:

$$\begin{aligned} Pr(\widetilde{\chi}_{ji}^{i,t} \mid s^{t-1}, a_i^{t-1}) = \beta Pr(\widetilde{\chi}_{ji}^{i,t-1} \mid s^{t-2}, a_i^{t-2}) \\ \sum_{a_j^{t-1}} \sum_{\widetilde{g}_j^{t-1}} Pr\left(a_j^{t-1} \mid s^{t-1}, \widetilde{\chi}_{ji}^{i,t-1}, \widetilde{g}_j^{t-1}\right) \\ \times T\left(s^{t-1}, a_i^{t-1}, a_j^{t-1}, s^t\right) Pr(\widetilde{\chi}_{ji}^{i,t} \mid \widetilde{\chi}_{ji}^{i,t-1}, a_j^{t-1}) \end{aligned} \quad (6)$$

where β is the normalizing constant and $Pr(\widetilde{\chi}_{ji}^{i,t} \mid \widetilde{\chi}_{ji}^{i,t-1}, a_j^{t-1})$ is the Kronecker delta function $\delta_K\left(\widetilde{a}_j^{t-1}, a_j^{t-1}\right)$. $\widetilde{a}_j^{t-1}$ is i's prediction of j's action given the estimated social intention $\widetilde{\chi}_{ji}^{i,t-1}$ and a_j^{t-1} is the actual action taken by j at the time step $(t-1)$. The Kronecker delta function evaluates to 1 only when the predicted action is the same as the actual action, thereby resolving Eq. 6 to:

$$\begin{aligned} Pr(\widetilde{\chi}_{ji}^{i,t} \mid s^{t-1}, a_i^{t-1}) = \beta Pr(\widetilde{\chi}_{ji}^{i,t-1} \mid s^{t-2}, a_i^{t-2}) \\ \sum_{\widetilde{g}_j^{t-1}} Pr\left(a_j^{t-1} \mid s^{t-1}, \widetilde{\chi}_{ji}^{i,t-1}, \widetilde{g}_j^{t-1}\right) \\ \times T\left(s^{t-1}, a_i^{t-1}, a_j^{t-1}, s^t\right) \end{aligned} \quad (7)$$

The social intention, estimated at time step t, is updated after actions taken by both the agents at each time step. This update is similar to the belief update in the POMDP framework but based on the estimated social intention policy of the other agent j.

E. Value Iteration

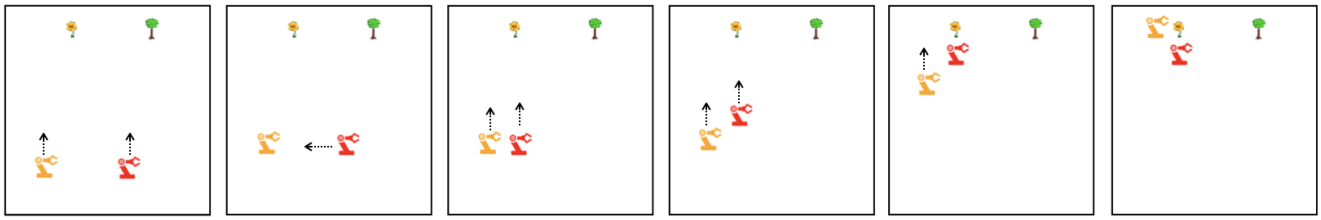
We use value iteration to solve S-MDP M_i^l for agent i at level l. The value function at $(k+1)$-th update of value iteration satisfies the following Bellman backup operation:

$$Q_i^{l,k+1}(s, a_i, a_j, \chi_{ij}) = R(s, a_i, \chi_{ij}) + \gamma \sum_{s' \in \mathcal{S}} T\left(s, a_i, a_j, s'\right) V_i^{l,k}\left(s', \chi_{ij}\right) \quad (8)$$

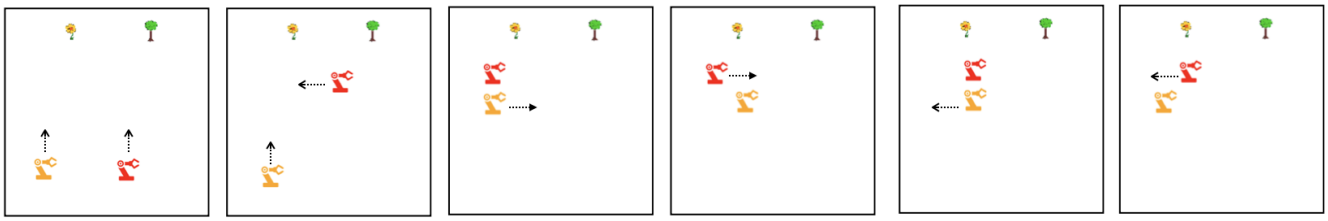
$$V_i^{l,k+1}(s, \chi_{ij}) = \max_{a_i \in \mathcal{A}_i} \left\{ \sum_{a_j \in \mathcal{A}_j} \int_{\widetilde{\chi}_{ji}^i} Pr(\tilde{\chi}_{ji}^i \mid s, a_i) \widetilde{\psi}_j^{i,l}(s', a_j', a_i', \widetilde{\chi}_{ji}^i) Q_i^{l,k+1}(s', a_i', a_j', \chi_{ij})\ d\widetilde{\chi}_{ji}^i \right\} \quad (9)$$

After applying Eq. 9 iteratively, agent i's optimal action for level l can be obtained as:

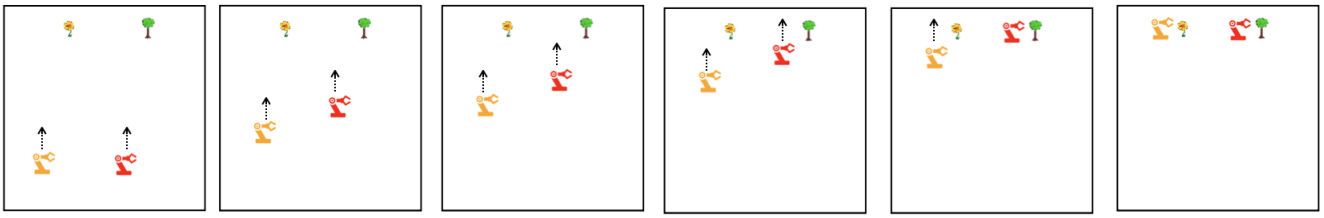
$$OPT(M_i^l) = \operatorname*{argmax}_{a_i \in \mathcal{A}_i} \left\{ \sum_{a_j \in \mathcal{A}_j} \int_{\widetilde{\chi}_{ji}^i} Pr(\tilde{\chi}_{ji}^i \mid s, a_i) \widetilde{\psi}_j^{i,l}(s', a_j', a_i', \widetilde{\chi}_{ji}^i) Q_i^{l,k+1}(s', a_i', a_j', \chi_{ij})\ d\widetilde{\chi}_{ji}^i \right\} \quad (10)$$



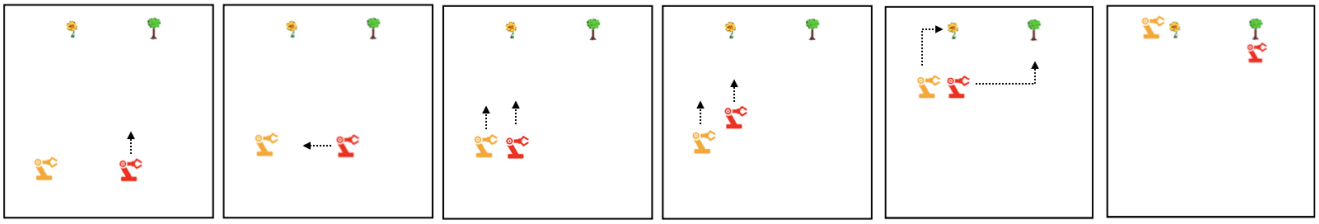
(a) Red robot is initialized with social intention of $\chi_{ji} = 1$ and shares the same goal with yellow robot of reaching the flower.



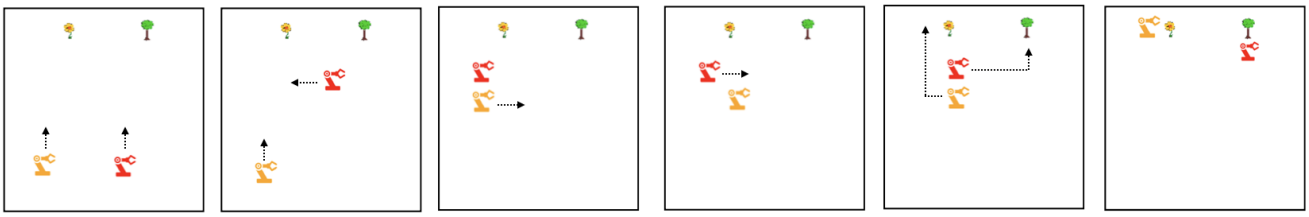
(b) Red robot is initialized with social intention of $\chi_{ji} = -1$ and shares the same goal with yellow robot of reaching the flower.



(c) Red robot is initialized with social intention of $\chi_{ji} = 0$ and has a different goal than yellow robot of reaching the tree.



(d) Red robot is initialized with social intention of $\chi_{ji} = 0.5$ and has a different goal than yellow robot of reaching the tree.



(e) Red robot is initialized with social intention of $\chi_{ji} = -0.5$ and has a different goal than yellow robot of reaching the tree.

Fig. 1: Example interactions between the red robot(agent j) and yellow robot(agent i). Red robot is initialized with different configurations of social intention towards yellow robot and physical goals.

F. Time Complexity

The time complexity of solving the S-MDP for agent i involves the cost of predicting the social intention policy $\widetilde{\psi}_j^{i,l-1}$ of the other agent at level l-1, which solves $\widetilde{\psi}_i^{j,l-2}$ and so on. Same as the typical MDP, the time needed for each iteration of the value iteration is $O\left(|\mathcal{A}_i||\mathcal{A}_j||\mathcal{S}|^2\right)$. At each level, the time taken to update the social intention is constant. The number of models at each level is then bounded by a number, $|\mathcal{M}|$, where $|\mathcal{M}|$ is the number of social intention χ evaluated at each level. Hence, solving the S-MDP is equivalent to recursively solving $O(|\mathcal{M}|^l)$ MDPs.

IV. EVALUATION

To show how the an agent can choose its behavior based on its estimations of the other agent, we apply our S-MDP framework to a multi-agent grid world inspired by previous studies on social perceptions [10, 4, 32]. It is a 7×7 2D grid

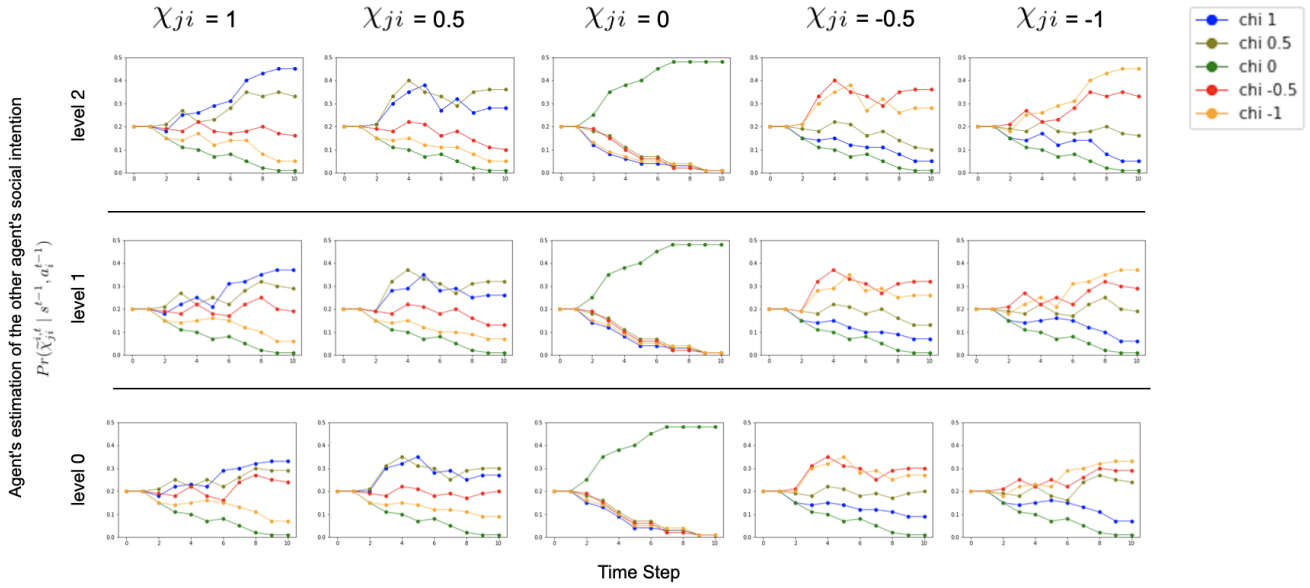


Fig. 2: Agent i's estimations of agent j's social intentions $\widetilde{\chi}_{ji}^{i}$ at different levels of reasoning and social intentions.

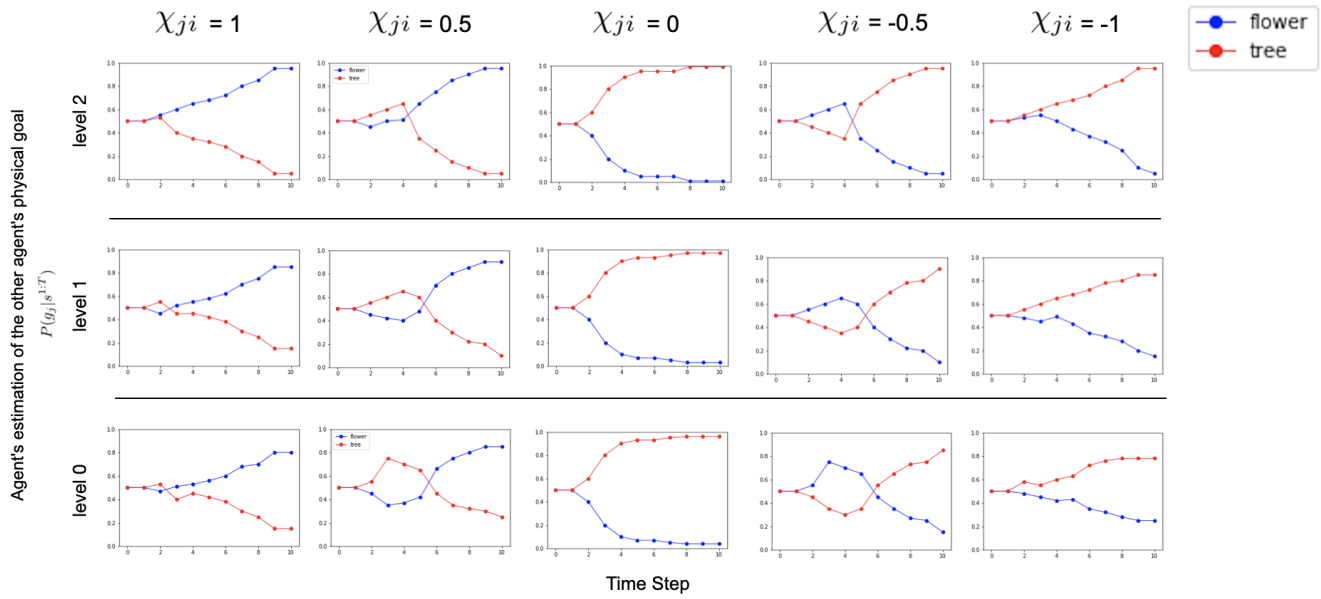


Fig. 3: Agent i's estimations of agent j's goal $\widetilde{g}_j$ at different levels of reasoning and social intentions.

consists of two agents, the yellow and red robots, and two physical landmarks, the flower and tree. The robots can have their physical goals (flower or tree), i.e. g_i and g_j to reach, and the social intention, χ_{ij} and χ_{ji} to help/hinder/ignore the other agent. The agents can move around the world by taking actions - *Left, Right, Up, Down* or choose to *Stay*. The agents reach their physical goal when they stay at the adjacent blocks of the landmarks.

In addition to social intentions, similar to [10], the agent's reward for reaching its physical goal is based on the agent's geodestic distance from the goal after taking an action. This physical reward function is parameterized by ρ and δ that determines the scale and shape of the physical reward: $r_i(s, a, g_i) = \max\left(\rho\left(1 - \text{distance}(s, a, g_i)/\delta\right), 0\right)$. We set the cost $c(a)$ of moving in the grid to 1 and staying at the same position to 0.1. The goal parameter ρ and δ were set to 1.25 and 5, respectively. The discount factor γ was set to 0.99. The value of social intention, χ_{ij} or χ_{ji}, varied between $[-1, 1]$ which corresponded from being most hindering to most helpful.

In this experiment, we use S-MDP to selection actions for the yellow agent (agent i) which has only a physical goal, reaching the flower or tree, and $\chi_{ij} = 0$ while interacting with the red agent (agent j). The red agent has a physical goal, reaching the tree, and social intention χ_{ji}. At every time step, the yellow agent estimates the social intention and the goal of the red agent at different levels of reasoning to predict its next action. In this evaluation, we run S-MDP in different scenarios by setting the groundtruth χ_{ji} to different values, -1, -0.5, 0, 0.5, and 1.

Figure 1 shows the sample interactions between the two agents at different scale of social intention χ_{ji}. When the red agent aggressively helps ($\chi_{ji} = 1$) the yellow, it goes directly to the yellow and stay together with the red. When the red agent politely hinders ($\chi_{ji} = -0.5$) the yellow, it goes to block the yellow's way to make the yellow inconvenient to reach the flower and then goes to its own goal. While $\chi_{ji} = 0$, both agents go their their respective goals directly.

We show the yellow agent's estimation of the red's social intention $\widetilde{\chi}^i_{ji}$ and physical goal $\widetilde{g}_j$ at different levels of reasoning and time steps in Figure 2 and 3. Each column is a scenario with a different groundtruth χ_{ji}. Each row shows the estimation when running S-MDP at level 0, 1, and 2. As the levels of reasoning increases, we find the estimated social intention and physical goal converges to the groundtruth more quickly in the aggressively-help/hinder scenarios. More levels also help the agent in the politely-help/hinder scenarios tell apart from the $\chi_{ji} = 1$ or $\chi_{ji} = -1$ hypothesis (see the difference between the blue and olive lines in $\chi_{ji} = -1$ and the difference between the red and orange lines in $\chi_{ji} = 1$ columns). The more accurate social intention estimation at earlier time steps also reflects in the amount of reward the yellow agent can collect. Figure 4 compares the cumulative reward at each time step for the yellow at different levels of reasoning and for the social intention $\chi_{ji} = 1$.

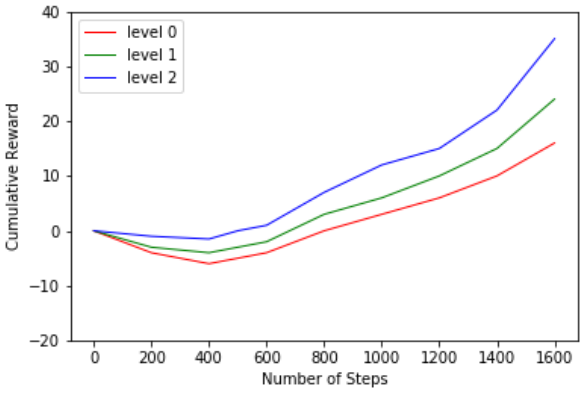


Fig. 4: Cumulative reward (the sum of all rewards received so far) of the agent i at different levels of reasoning and for the social intention $\chi_{ji} = 1$

V. CONCLUSION & FUTURE WORK

We presented S-MDPs to efficiently incorporate social interactions into MDP framework. We achieve this by (1) recursively estimating the social intentions and the goals of other agents and (2) extending the reward function to include the estimates of other agents (by maximizing or minimizing the estimated reward of the other agent). S-MDPs enable zero-shot social inference of other agents. An initial experiment in a multi-agent 2D grid world showed that such multiple levels of reasoning improves the estimations of the other agent's social intention and physical goal as well as the accumulated rewards. This type of social agents need extra computations as it solves multiple MDPs recursively. The increased complexity is bounded by the levels of reasoning needed to model social interactions. In the future, we plan to investigate how the learned policies changes in different levels/scenarios and validate with human experiment to understand what number of levels of recursive reasoning needed and what features are useful to model social interactions with S-MDPs. While we only present S-MDP here modeling the recursive social rewards, it is possible to extend it with I-POMDP to cover more space of social interactions to further enable rich human-robot interactions.

REFERENCES

[1] R. Bellman, "A markovian decision process," *Journal of mathematics and mechanics*, vol. 6, no. 5, pp. 679–684, 1957.

[2] K. J. Åström, "Optimal control of markov processes with incomplete state information," *Journal of Mathematical Analysis and Applications*, vol. 10, no. 1, pp. 174–205, 1965.

[3] P. J. Gmytrasiewicz and P. Doshi, "Interactive pomdps: Properties and preliminary results," in *International Conference on Autonomous Agents: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-*, vol. 3, 2004, pp. 1374–1375.

[4] C. L. Baker, N. D. Goodman, and J. B. Tenenbaum, "Theory-based social goal inference," in *Proceedings of the thirtieth annual conference of the cognitive science society*. Cognitive Science Society, 2008, pp. 1447–1452.

[5] C. L. Baker and J. B. Tenenbaum, "Modeling human plan recognition using bayesian theory of mind," *Plan, activity, and intent recognition: Theory and practice*, vol. 7, pp. 177–204, 2014.

[6] J. Kiley Hamlin, T. Ullman, J. Tenenbaum, N. Goodman, and C. Baker, "The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model," *Developmental science*, vol. 16, no. 2, pp. 209–226, 2013.

[7] M. Kleiman-Weiner, M. K. Ho, J. L. Austerweil, M. L. Littman, and J. B. Tenenbaum, "Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction," in *CogSci*, 2016.

[8] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. A. Eslami, and M. Botvinick, "Machine theory of mind," in *International conference on machine learning*, 2018, pp. 4218–4227.

[9] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[10] T. D. Ullman, C. L. Baker, O. Macindoe, O. Evans, N. D. Goodman, and J. B. Tenenbaum, "Help or hinder: Bayesian models of social goal inference," Tech. Rep., 2009.

[11] X. Puig, T. Shu, S. Li, Z. Wang, J. B. Tenenbaum, S. Fidler, and A. Torralba, "Watch-and-help: A challenge for social perception and human-ai collaboration," *arXiv preprint arXiv:2010.09890*, 2020.

[12] S. V. Albrecht and P. Stone, "Autonomous agents modelling other agents: A comprehensive survey and open problems," *Artificial Intelligence*, vol. 258, pp. 66–95, 2018.

[13] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "Cooperative inverse reinforcement learning," *arXiv preprint arXiv:1606.03137*, 2016.

[14] R. DeVries, "Piaget's social theory," *Educational researcher*, vol. 26, no. 2, pp. 4–17, 1997.

[15] G. P. Dimuro, A. C. da Rocha Costa, L. V. Gonçalves, and D. R. Pereira, "Recognizing and learning models of social exchange strategies for the regulation of social interactions in open agent societies," *Journal of the Brazilian Computer Society*, vol. 17, no. 3, pp. 143–161, 2011.

[16] L. F. Macedo, G. P. Dimuro, M. S. Aguiar, A. C. Costa, V. L. Mattos, and H. Coelho, "Analyzing the evolution of social exchange strategies in social preference-based mas through an evolutionary spatial approach of the ultimatum game," in *2012 Third Brazilian Workshop on Social Simulation*. IEEE, 2012, pp. 83–90.

[17] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in tv shows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, 2012.

[18] M. J. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari, "Detecting people looking at each other in videos," *International Journal of Computer Vision*, vol. 106, no. 3, pp. 282–296, 2014.

[19] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 1593–1600.

[20] F. Heider and M. Simmel, "An experimental study of apparent behavior," *The American journal of psychology*, vol. 57, no. 2, pp. 243–259, 1944.

[21] T. Shu, M. Kryven, T. D. Ullman, and J. B. Tenenbaum, "Adventures in flatland: Perceiving social interactions under physical dynamics," in *42d proceedings of the annual meeting of the cognitive science society*, 2020.

[22] M. Kryven, T. Ullman, B. Cowan, and J. Tenenbaum, "Plans or outcomes: How do we attribute intelligence to others?" 2021.

[23] A. Netanyahu, T. Shu, B. Katz, A. Barbu, and J. B. Tenenbaum, "Phase: Physically-grounded abstract social events for machine social perception," *arXiv preprint arXiv:2103.01933*, 2021.

[24] P. Doshi and P. J. Gmytrasiewicz, "Monte carlo sampling methods for approximating interactive pomdps," *Journal of Artificial Intelligence Research*, vol. 34, pp. 297–337, 2009.

[25] P. Doshi and D. Perez, "Generalized point based value iteration for interactive pomdps." in *AAAI*, 2008, pp. 63–68.

[26] B. Ng, K. Boakye, C. Meyers, and A. Wang, "Bayes-adaptive interactive pomdps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012.

[27] P. Doshi, X. Qu, A. Goodie, and D. Young, "Modeling recursive reasoning by humans using empirically informed interactive pomdps," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, 2010, pp. 1223–1230.

[28] C. F. Camerer, T.-H. Ho, and J.-K. Chong, "A cognitive hierarchy model of games," *The Quarterly Journal of Economics*, vol. 119, no. 3, pp. 861–898, 2004.

[29] I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins, "Introduction to nested markov models," *Behaviormetrika*, vol. 41, no. 1, pp. 3–39, 2014.

[30] W. Yoshida, R. J. Dolan, and K. J. Friston, "Game theory of mind," *PLoS Comput Biol*, vol. 4, no. 12, p. e1000254, 2008.

[31] T. N. Hoang and K. H. Low, "Interactive pomdp lite: Towards practical planning to predict and exploit intentions for interacting with self-interested agents," *arXiv preprint arXiv:1304.5159*, 2013.

[32] C. Baker, R. Saxe, and J. Tenenbaum, "Bayesian theory of mind: Modeling joint belief-desire attribution," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, no. 33, 2011.