Social Interactions as Recursive MDPs

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

INTRODUCTION

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. 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.

CONTRIBUTIONS

- 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.
- An implementation where that function is a linear transformation, which captures notions of helping and hindering
- Experimental validation of zero-shot social understanding where agents that have never been asked to help or hinder do so.

S-MDP

S-MDP for an agent *i* at level *l* is defined as:

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

where χ_{ij} represents the social intention of agent *i* towards agent *j* and is used in reward function to define the reward in helping or hindering the other agent *j*.

Reward
Each agent can have its own physical goal, e.g. going to a landmark, as well as the social goals(intention), 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:

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

where r(.) is the static reward given the agent's own physical goal (e.g. *g_i* and *g_j*) and c(.) is the cost for taking an action.

Estimating goals and social intentions
Similar to Shu et al. (2020) , the physical goal *g_j* of agent *j* is predicted by *i* using the Bayes's rule:

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

The social intention of agent *i* towards agent *j* estimated by agent *k* at level *l* is denoted as $\widetilde{\chi}^{k,t}_{ij}$ (described below in social intention update). In the two-player setting, *k* can be either agent *i* or *j* depending on which agent is making estimation.

Planning for S-MDP
S-MDP considers the expectation over the estimated social intention of agent *j* in the Q function:

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

The *l*-level social intention policy of the agent *j* is predicted by *i* using the Q function at level *l-1* as:

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

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..,l-1*.

Social Intention Update
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:

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

where β is the normalizing constant and $Pr(\widetilde{\chi}^{i,t}_{ji} \mid \widetilde{\chi}^{i,t-1}_{ji}, a_j^{t-1})$ is the Kronecker delta function $\delta_\kappa(\widetilde{a}_j^{t-1}, a_j^{t-1})$. $\widetilde{a}_j^{t-1}$ is *i*'s prediction of *j*'s action given the estimated social intention $\widetilde{\chi}^{i,t-1}_{ji}$ 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:

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

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*.

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:

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

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

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

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

EVALUATION

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.

Red robot is initialized with social intention 1 and shares the same goal with yellow robot of reaching the flower.

Red robot is initialized with social intention of -1 and shares the same goal with yellow robot of reaching the flower.

Red robot is initialized with social intention of 0.5 and has a different goal than yellow robot of reaching the tree.

Red robot is initialized with social intention of -0.5 and has a different goal than yellow robot of reaching the tree.

Yellow agent's estimation of the red's social intention and physical goal at different levels of reasoning and time steps in shown below: