Unveiling the latent dynamics in social cognition with multi-agent inverse reinforcement learning

Yusi Chen*
Computational Neuroscience Center
University of Washington
cyusi@uw.edu

Angela Radulescu
Department of Neuroscience and Friedman Brain Institute
Icahn School of Medicine at Mount Sinai
angela.radulescu@mssm.edu

Herbert Zheng Wu†
Department of Neuroscience and Friedman Brain Institute
Icahn School of Medicine at Mount Sinai
zheng.wu@mssm.edu

Abstract

Understanding the intentions and beliefs of others, a phenomenon known as "theory of mind", is a crucial element in social behavior. These beliefs and perceptions are inherently subjective and latent, making them often unobservable for investigation. Social interactions further complicate the matter, as multiple agents can engage in recursive reasoning about each other's strategies with increasing levels of cognitive hierarchy. While previous research has shown promise in understanding a single agent's belief of values through inverse reinforcement learning, extending this to model interactions among multiple agents remains an open challenge due to the computational complexity. In this work, **we adopted a probabilistic recursive modeling of cognitive levels and joint value decomposition to achieve efficient multi-agent inverse reinforcement learning (MAIRL)**. We validated our method using simulations of a cooperative foraging task. Our algorithm revealed both the ground truth goal-directed value function and agents' beliefs about their counterparts' strategies. When applied to human behavior in a cooperative hallway task, our method identified meaningful goal maps that evolved with task proficiency and an interaction map that is related to key states in the task without accessing to the task rules. Similarly, in a non-cooperative task performed by monkeys, we identified mutual predictions that correlated with the animals' social hierarchy, highlighting the behavioral relevance of the latent beliefs we uncovered. Together, our findings demonstrate that MAIRL offers a new framework for uncovering human or animal beliefs in social behavior, thereby illuminating previously opaque aspects of social cognition.

*Correspondence
†Correspondence

Preprint. Under review.

1 Introduction

Humans are remarkable at understanding the mental states of others - their beliefs, values, and intentions - a capacity known as "theory of mind (ToM)." At its core, an individual's behavior is often not directly guided by objective reality, but by their subjective beliefs about the state of the world. "Theory of mind" thus allows us to more reliably predict other people's actions instead of using the physical reality or one's own belief [11]. This reasoning process can operate recursively, with individuals modeling not just others' beliefs, but others' beliefs about their own beliefs - a phenomenon captured by increasing levels of cognitive hierarchy. Precisely inferring these often unobservable beliefs and subjective values is crucial for understanding social cognition in humans as well as other species. We adopt a multi-agent inverse reinforcement learning (MAIRL) framework to achieve such inference. Addressing a crucial challenge in neuroscience and cognitive science, our novel approach promises to reveal the latent model of the world and the computation of value and mental states of multiple interacting agents in social behavior [12].

Multi-agent reinforcement learning (MARL) models interactions among multiple agents [28][7]. Despite its wide application in topics including traffic coordination [8], wireless sensor networks [10], and patient coordination in healthcare [20], its application in neuroscience and cognitive sciences has been limited. There are many categories of MARL algorithms but we focus mainly on those with centralized training and different levels of execution, in which agents share a common critic, but execute policies independently. A review of relevant algorithms can be found in this benchmark paper [18].

In many social interactions, an agent anticipates the behaviors of others by constructing predictive models of them. This essential aspect of interaction is captured by recursive probabilistic reasoning, often used in the field of opponent modeling (OM) [9][2][23]. However, OM is often not incorporated into modern RL frameworks due to computational expense and the need for centralized training to allow access to the ground truth strategies by all agents [26]. Nevertheless, in well-defined social tasks, participants are expected to fully understand the rules, objectives, and their opponents' strategies. This understanding justifies the use of recursive modeling and the assumption that participants will make the best responses based on their predictions. Additionally, recent approaches have aimed to decouple agents' mutual effects during training through variational Bayes methods, enabling the approximation of task solutions [26] and opening more possibilities for large-scale OM.

While MARL with recursive reasoning models social behavior using known value functions, real-world interactions among humans or animals often lack explicitly defined value functions. Inverse reinforcement learning (IRL) addresses this by inferring the latent value functions from observed behaviors [3], thereby uncovering an individual's model of the world. Assuming the value function is a linear combination of known feature functions, the problem can be formulated as a linear optimization to maximize expert feature expectations, with various approaches addressing computation complexity and solution ambiguities [1][19][31]. We focused on the maximum entropy approach proposed in [31], as it provides a probabilistic version of the problem, facilitating the incorporation of recursive reasoning crucial for ToM modeling.

Merging these lines of work, our approach uncovers both goal related value functions of the agents and mental models of one agent for another. In Section 2, we first review related work that applies IRL, or more generally, Bayesian inference, to model social interactions in neuroscience and cognitive sciences. In Section 3, we describe our formulation based on Markov Decision Processes (MDP) with recursive reasoning and value decomposition to model two expert agents in simultaneous-move games and a maximum a posteriori algorithm for inference. In Section 4, we show that our model **effectively recovers goal-related value functions, interaction terms associated with task rules, and the levels of cognitive hierarchy between two agents** in a simulated multi-state, multi-action cooperative foraging task. We further apply our model to experimental datasets in humans and non-human primates, uncovering previously unobservable value maps and latent strategies used by the players.

2 Related work

Most popular approaches to studying social interactions falls under the umbrella of Bayesian inference as reviewed in Jara-Ettinger [14]. Based on Partially Observable Markov Decision Processes

(POMDP), most work focuses on one agent's observable behavior and hidden mental state [6][5]. While fruitful, this approach has several drawbacks. First, there is a lack of consistent mapping between the inferred value function and the specific mental state of the agent such as goal [5], desire [6], context-specific intentions [25] or intensity of the behavior [21]. Secondly, most research works under single state setup [21][16] or binary intent inference [24] because extending the POMDP approach to the multiagent, multi-state scenario is computational challenging [27].

Alternatively, based on Markov Decision Process (MDP), Yoshida et al. [29] took a different approach by inferring the opponents' ToM for optimal decision-making. Their work optimized the energy function of a given policy, thus representing one special case of our work. The inferred goal map and levels of cognitive hierarchy remain entangled because they inferred different values functions at each level.

To address these gaps and challenges in the previous work, we adopt **a multi-agent MDP with value decomposition and recursive reasoning** to enable the inference of each agent's goal related value functions and internal predictions in social interactions.

3 MAIRL: multi-agent inverse reinforcement learning

3.1 Problem formulation

Let's consider a Markov Decision Process [3] described by $\langle \mathcal{S}, \mathcal{A}, P, r^1, r^2, \gamma \rangle$ where:

- $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ is the set of joint environmental states constructed by the cartesian product of individual state sets $\mathcal{S}_i$ with cardinality $|\mathcal{S}| = |\mathcal{S}_1| \times |\mathcal{S}_2|$
- $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ is the set of joint actions
- $P = P(s'|s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ describes the state transition dynamics of the environment, where Δ is a probability measure over $\mathcal{S}$.
- $r^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the reward function that returns a scalar value to the i-th agent for a transition from $s \in \mathcal{S}$, taking joint action $a \in \mathcal{A}$ to, $s' \in \mathcal{S}$.
- $\gamma \in [0, 1]$ is the discount factor for future steps.

The inverse reinforcement learning (IRL) problem [1][31][4] states that: given $\{P, \mathcal{S}, \mathcal{A}, \gamma\}$ and N samples of expert trajectories $D = \{\zeta_1, \zeta_2, ..\zeta_N\}$, we aim to infer the unknown reward function r^1, r^2 in such a way that $P(D|r^1, r^2)$ is maximized. Each trajectory is composed of independent state-action pairs: $\zeta_i = \{(s_t, a_t)\}_{t=0}^{T}$. As a consequence, the posterior probability of observing expert trajectories ζ_i can be calculated with the specific policy π derived from value functions (Eq. 1).

$$P(\zeta_i|r) = \prod_{t=0}^{T} \pi(a_t|s_t)p(s_t) \tag{1}$$

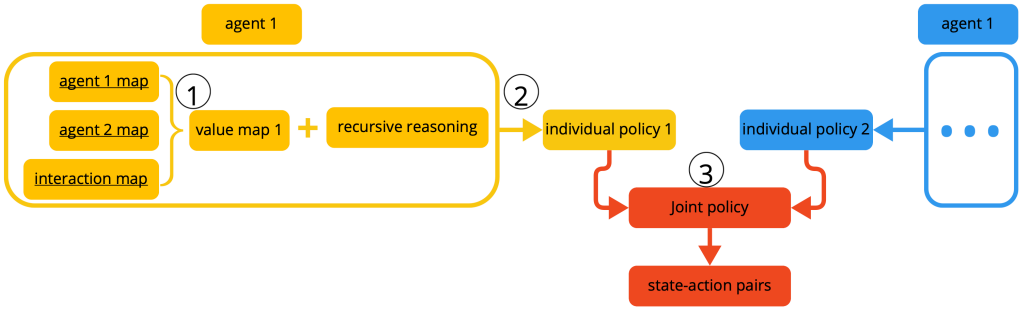
Therefore, at the core of this problem is to parameterize the *joint policy function* $\pi(a|s)$ using a probabilistic modeling of social interactions based on MARL with recursive reasoning and value decomposition as highlighted in Figure 1.

3.2 Value function parameterization and decomposition

Previous efforts [4] to infer goal maps in animal foraging parameterized the reward function through a low rank representation of goal maps: $r(s) = \sum_{k=1}^{K} \alpha_k u_k(s)$, which leads to $K(1 + |\mathcal{S}|)$ number of parameters to infer. Such a formulation is computationally expensive in a multi-agent scenario, as the cardinality of $\mathcal{S}$ grows exponentially with the number of agents. To alleviate the issue, we adopted value decomposition such that a joint map could be decomposed into marginal maps and interactions maps as Eq. 2 where $s_i \in \mathcal{S}_i$ and ϕ stands for an given interaction function between s_1 and s_2.

$$u(s_1, s_2) = m(s_1) + n(s_2) + \beta\phi(\|s_1 - s_2\|) \tag{2}$$

[3] For simplicity, we provide equations for two agents. These derivations naturally extend to many agents albeit with exponential growth of required computation.



$$\alpha, m, n, \phi \xrightarrow{①} r^i \xrightarrow{②} \pi^i \xrightarrow{③} \pi \rightarrow (s_t, a_t) \in D$$

Figure 1: The forward generation graph of making social decisions. We observe state-space pairs and attempts to infer the emphasized (underlined) components of the decomposed value map per agent. ①: value decomposition (Eq. 2), ②: policy generation with soft Q-learning (Eq. 3 and Eq. 4) and recursive reasoning (Table 1), ③: centralized or decentralized policy execution.

In the case of cooperative open arena foraging, we showed that the joint reward could be perfectly reconstructed using individual maps (m, n) and an interaction map (ϕ) that takes into account the Euclidean location distance between agents (Fig. S1 and Supplementary).

Therefore, our objective is to learn the map weights α_k , individual maps $m(s_1)$, $n(s_1)$ and interactions maps $\phi(s_1, s_2)$ that maximize the posterior of observing expert trajectories $D = \{(s_t, a_t)\}$.

3.3 Policy parameterization

To facilitate the inference problem, we adopted a differentiable maximum entropy policy formulation [31][32]:

$$\pi(a|s) = \frac{\exp Q(s,a)}{\sum_{a' \in \mathcal{A}} \exp Q(s,a')} \tag{3}$$

where $Q(s, a)$ is a soft Q-function arising from performing soft value iteration:

$$Q(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) \log(\sum_{a' \in \mathcal{A}} \exp Q(s,a')) \tag{4}$$

Note that a temperature term in not needed in the softmax function, as it will be absorbed into absolute value of Q, and subsequently r, which is the variable we aim to estimate.

In the multi-agent scenario, the above generative model only holds for a cooperative task where two agents have the same reward function and act in a centralized way. For non-collaborative tasks, we need separate r^i, Q^i and $\pi^i(a|s)$ for each agent. Other cases involving decentralized policy execution will be discussed in the next sections.

3.4 Levels of cognitive hierarchy

The major complexity of modeling multi-agent interactions is the level of coordination and mutual predictions between agents. For notation simplicity, let's consider collaborative tasks where each agents have an identical value function. We further assume that agents know the state information of each other. A summary of the cognitive levels considered in this work was listed in Table 1.

In one extreme, agents generate an optimal joint policy $\pi(a_1, a_2|s)$ according to Eq. 3 and operate in a centralized way such that the optimal **joint action** would be executed. In this case, standard IRL formulations in Eq. 3 and 4 could be applied to calculate the posterior probability in an exponentially enlarged state space and action space. Note that in the following context, π refers specifically to the optimal joint policy, not necessarily the policy adopted by agents.

Table 1: Generative models with different coordination levels.

Generative model	Prediction model	Joint action probability
Centralized	/	$\pi(a_1, a_2\|s)$
Independent control	egocentric agents	$P(a_1\|s_1)P(a_2\|s_2)$
Independent control	chance prediction	$\prod_i \|\mathcal{A}_{-i}\|^{-1} \sum_{a_{-i}} \pi(a_i\|a_{-i}, s)$
Independent control	optimal prediction	$\pi(a_1\|s)\pi(a_2\|s)$

In a less coordinated scenario, agents still generate an optimal joint policy π but the policy execution process is independent, which is often the case in simultaneous action games. Mathematically, we have conditional independence of individual policies:

$$P(a|s) = P(a_1|s)P(a_2|s) \tag{5}$$

Then the question becomes how does each agent chooses its own action based on full state information. The answer lies in a probabilistic view of recursive reasoning.

For a level-0 or egocentric agent, who does not consider the action or state of other agents, its policy only depends on its own state information, i.e. $P(a_i|s) = P(a_i|s_i)$ and could be generated as soft value iteration of its own map function (m or n).

For an agent with level-1 reasoning or higher, it builds a predictive model of the other player and is expected to act optimally based on this prediction. Let's denote agent 1's prediction of agent 2's policy as $\hat{a}_2$, based on which, we could break the policy of agent 1 as:

$$P(a_1|s) = \sum_i P(a_1, \hat{a}_2 = i|s) = \sum_i P(a_1|\hat{a}_2 = i, s)P(\hat{a}_2 = i|s) \tag{6}$$

Since a rationale agent 1 wants to act optimally for the joint task based on its own predictions, the first term is essentially the conditioned optimal policy:

$$P(a_1|\hat{a}_2, s) = \pi(a_1|a_2, s) \tag{7}$$

Thus the agent's prediction about its counterpart's policy as well as its own joint value map determines the formulation of its policy. The recursive level of this prediction also determines the agent's level of reasoning. For example, if agent 1 assumes agent 2 to be an egocentric, or level-0 agent (regardless of whether agent 2 is or not), then agent 1 has a cognitive level of one. If agent 2 assumes agent 1 is level-1 and forms its prediction about agent 1's policy (i.e. $P(\hat{a_1}|s)$) based on this assumption, agent 2 has a cognitive level of two. Even higher levels of reasoning can be constructed using this recursive process. Since we are interested in simple social tasks and even human reasoning has a depth of one or two levels [15], we limit our consideration here to two simple and extreme parameterizations of mutual prediction that provide the lower and upper bounds in terms of optimality in the case of independent control with prediction:

- If agent 1 has no information about agent 2's policy and generates chance level prediction (i.e. $P(\hat{a}_2|s) = |\mathcal{A}_2|^{-1}$), then $P(a_1|s) = |\mathcal{A}_2|^{-1} \sum_{a_2} \pi(a_1|a_2, s)$.
- If agent 1 predicts that agent 2 would act on the optimal marginal policy (i.e. $P(\hat{a}_2|s) = \pi(a_2|s)$), then $P(a_1|s) = \sum_i P(a_1|a_2 = i, s)\pi(a_2 = i|s) = \pi(a_1|s)$. In other words, agent 1 also acts on the marginal probability of the optimal joint policy.

3.5 Inference procedure

The inference procedure is adapted from Ashwood et al. [4] except that we have time-invariant version of weight estimation and different evaluation of joint policy given a chosen generative model for cognitive levels. Let's assume $K = 1$, our objective is to learn the map weights $\boldsymbol{\alpha}$, individual maps $m(s_1)$, $n(s_1)$ and interactions maps $\phi(s_1, s_2)$ that maximize the posterior of observing expert trajectories $D = \{(s_t, a_t)\}$ under different generative models listed in Table 1.

We assume all parameters have a Gaussian prior with known variance and mean and individual maps have different prior variance compared to the interaction map. Mathematically, $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{1}, \Sigma)$ where diag$(\Sigma) = (\sigma_0^2, \sigma_0^2, \sigma_1^2)$, $m, n \sim \mathcal{N}(\mathbf{0}, \Sigma_m)$, $\phi \sim \mathcal{N}(\mathbf{0}, \Sigma_\phi)$. Note that we assume Σ_m and Σ_ϕ are diagonal matrices. Incorporating the map prior is equivalent to adding an L-2 regularizer with

Algorithm 1 MAIRL with value decomposition

Input1: MDP information: $\langle \mathcal{S}, \mathcal{A}, P, \gamma \rangle$;
Input2: Expert trajectories D;
Input3: Hyperparameters: K, σ_0, σ_1, λ_1, λ_2; the chosen generative model in Table 1.;
Initialize $\boldsymbol{\alpha}, m, n, \phi$;
for i = 1...N_{iter} **do**
  Calculate rewards r as a function of $\boldsymbol{\alpha}, m, n, \phi$ (Eq. 2);
  Get *optimal joint policy* π using softmax iteration until convergence (Eq. 3 and 4);
  Get *policy* $P(a|s)$ based on the chosen generative model of cognitive level (Table 1);
  Update $\boldsymbol{\alpha}$ given reward map fixed (Eq.8, line 1);
  Calculate rewards r and generate new policy as before;
  Update m, n, ϕ given weights fixed (Eq.8, line 2);
end for
Output: $\boldsymbol{\alpha}, m, n, \phi$

coefficients λ_1 and λ_2 to the map entries. Mathematically, the parameter optimization progress could be written as:

$$\begin{aligned} \boldsymbol{\alpha}^* &= \text{argmax}_{\alpha} \sum_{(s_t,a_t)\in D} \log P(s_t, a_t) - \frac{1}{2}((\boldsymbol{\alpha} - \mathbf{1})^T \Sigma^{-1} (\boldsymbol{\alpha} - \mathbf{1})) \\ (m^*, n^*, \phi^*) &= \text{argmax}_{m,n,\phi} \sum_{(s_t,a_t)\in D} \log P(s_t, a_t) - \lambda_1 \|m\|^2 - \lambda_1 \|n\|^2 - \lambda_2 \|\phi\|^2 \end{aligned} \tag{8}$$

We performed coordinate ascent to iteratively update weights and maps while holding the other set of parameters fixed as illustrated in Alg. 1.

4 Results

4.1 Validation using a simulated cooperative foraging task

We first validated our method using simulations of a cooperative foraging task. In this task, two artificial agents navigate freely in an open arena of a five-by-five grid. On every trial, the two artificial agents are presented with two possible reward locations, but are only rewarded if they arrive simultaneously at the same target location (yellow squares in Fig. 2A). Both agents share the same ground truth reward function, which is defined over the joint state space and has non-zero values only at the states where both agents are at the reward locations. Simulation and fitting details can be found in the Supplementary.

4.1.1 Infer marginalized value maps and task dependent interaction map

The paths of two skilled agents (Fig. 2B and C) were simulated and analyzed using MAIRL with centralized learning and control. The resulting value maps perceived by both agents are shown in Fig.2D, E, and F. The target locations have higher value in the individual marginal goal maps, representing the available reward at these locations. Further, the map of Euclidean distance between the two agents (Fig. 2F) peaks when the agents are close to each other. This faithfully captures the key social element for solving this task: staying close to the partner. The larger ϕ compared to m and n also indicates that the agents place a higher subjective value on social proximity than being at the reward location. Importantly, the same inference algorithm without value decomposition resulted in a significantly lower data likelihood, likely getting stuck in a local minimum due to the excessive number of parameters to estimate (Fig. 2G). Thus, our MAIRL algorithm with value decomposition reveals the precise value each agents assigns to the two rewarding features in this tasks and invites further investigation of the neural representation of these values.

4.1.2 Identify levels of cognitive hierarchy by model comparison

Next, we infer the levels of cognitive hierarchy for the two agents using the simulated trajectories. The task lengths, represented by the number of foraging steps required to obtain rewards, appeared

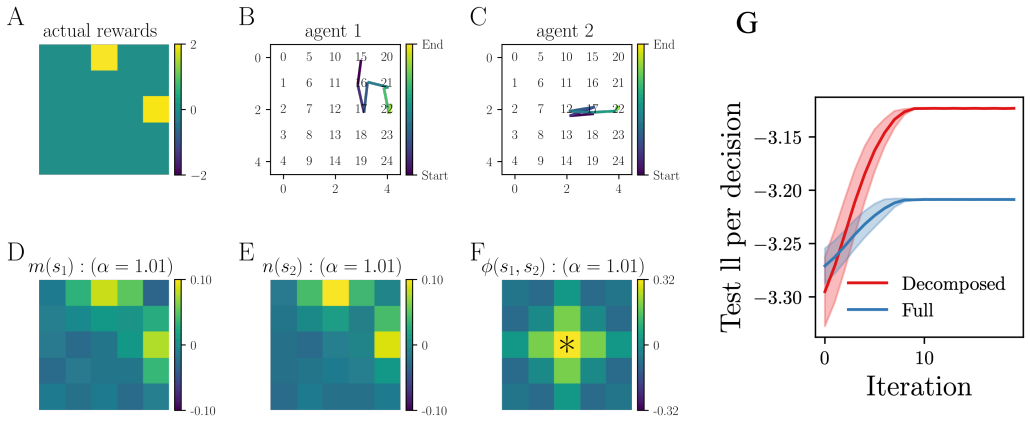


Figure 2: **Inference of decomposed value maps in a cooperative foraging task.** (A) In a five-by-five open arena, two artificial agents are only rewarded if they simultaneously arrive at the target locations (yellow cells). (B) (C) Example foraging trajectories of two expert agents starting from random positions and arriving at the east target location simultaneously. (D) (E) Estimated value map and weights (α) of agent 1 or agent 2's position in the arena. (F) Estimated value map of social proximity (ϕ). The heatmap indicates the values of one agent's position relative to the other who is placed at the origin for plotting purpose and marked with an asterisk. (G) Log likelihood (ll) of the test dataset along the iteration process of MAIRL with or without value decomposition. Each iteration step involves 40 full batch gradient descent steps.

indistinguishable across models (Fig. S2). We compared models using Akaike Information Criterion (AIC) of the test trajectories. Conventionally, a drop of 2 or more in AIC is considered statistically significant. As shown in Fig. 3A, MAIRL effectively identified the ground truth model, except for close scores of the two independent control models when fitting the simulated chance prediction data (see limitations in Discussion). These findings underscore the efficacy of our approach in accurately inferring cognitive hierarchy from similar experimental trajectories.

In another simulation, where an egocentric agent was paired with a thoughtful level-1 agent, the task was still accomplished with a trajectory indistinguishable from that of a centralized pair (Fig. S2). By fitting MAIRL to the individual trajectories, we were able to identify the egocentric agent (Fig. 3B), but not the level-1 agent (Fig. 3C). This may be because this simple task does not require a sophisticated mentalization process.

4.2 Applications to experimental datasets

After validating in simulations where ground truth is available, we extended our approach to analyze two real datasets. The first dataset is collected from human participants performing a cooperative hallway task, and the second dataset is from monkeys performing a non-cooperative "chicken" game. By applying MAIRL, we recovered goal related value maps, task rule related interaction maps and the level of cognitive hierarchy, offering novel insights into the participants' subjective strategies and their social dynamics.

4.2.1 Infer human participants' strategy and task rules in a cooperative task

The hallway task [13] requires participants to start from orange or blue circles and switch locations without colliding. A total of 397 trials of human trajectories were collected and categorized into three groups: fail, success, and expert (completed with minimal number of steps), indicating increasing skill levels. Example trajectories are shown in Fig. S4.

We applied models of different cognitive levels to fit these trajectories (Fig. 4B). Pairs from expert trials exhibited remarkable coordination, akin to centralized control, despite the fact that each player was acting independently. This was evident from their capability for mutual predictions, as the value map of one agent is accurately predicted by the other (Fig. 4C). Pairs in successful trials displayed

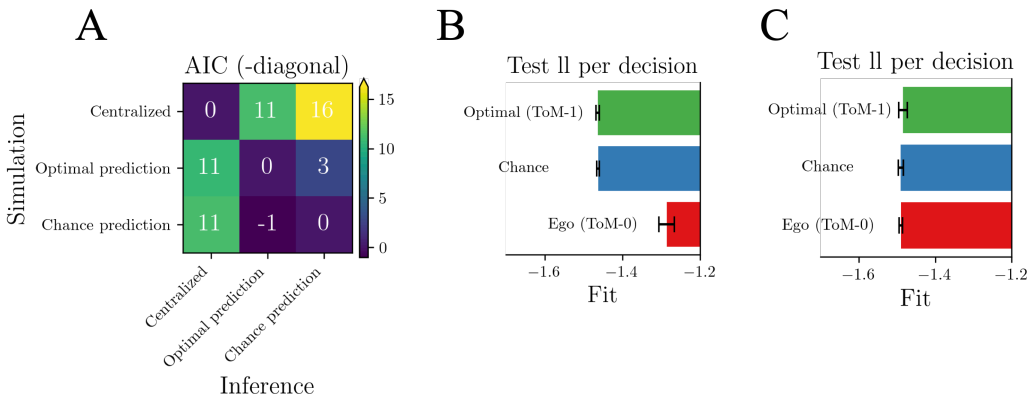


Figure 3: **Inference of cognitive hierarchy from similar trajectories.** (A) Akaike Information Criterion (AIC) of the test set simulated using the models listed on the vertical axis and fitted with the models listed on the horizontal axis. A decrease of 2 or more in AIC indicates a significantly better fit for the model. (B)(C) Log likelihood (LL) (mean $\pm$ std) of the test trajectories simulated for a pair of an egocentric agent and a level-1 agent. Trajectories of the egocentric individual (B) or level-1 agent (C) were fitted using generative models listed on the vertical axis.

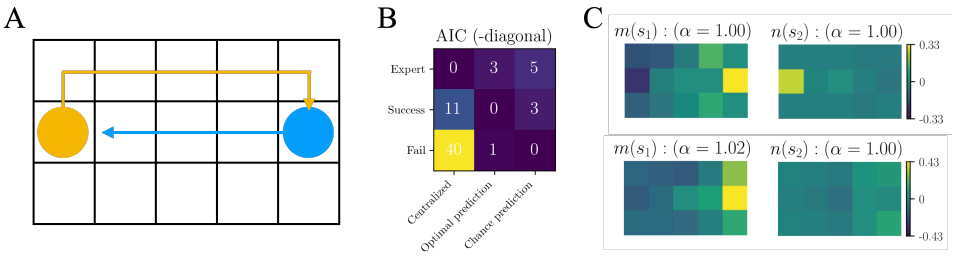


Figure 4: **Human participants' develop good estimation of each other to succeed for collaboration.** (A) Setup of the hallway task (B) Model fitness by AIC for trajectories of the three categories: fail, success and expert. (C) Estimated value maps of agent 1's trajectories from expert pairs (top) and failed pairs (bottom) using an optimal prediction model. In fail trials, the participant did not attach value to its counterpart's location (s_2). The value maps are similar when fitted using a chance prediction model (Fig. S3).

less coordination but maintained good mutual predictive capacity, supported by a well-fitted model of optimal prediction. For pairs in failed trials, trajectories were similarly explained by both optimal and chance prediction models. Nonetheless, inferred value maps using both models indicated that both participants did not consider each other's location (Fig. 4C and Fig. S3).

Unlike in cooperative foraging, where the key is to stay close to each other, the hallway task does not clearly indicate using Euclidean distance as an interaction term. Therefore, we tested different interaction maps, including Euclidean distance, column distance, row distance, and a joint map of both. In both expert and failed pairs, the model that used column distance as the interaction term performed best (Fig.5A). The importance of column distance highlights the essence of the hallway task: switching column indices without colliding. Thus, states where the column distance equals zero (Fig.5C) are critical to solving the task. The recovered values showed that expert pairs highly valued these key states (Fig.5B). In contrast, failed pairs often collided and were bounced back to states where the column distance equaled one (Fig.5D), leading them to assign higher value to these states related to colliding (Fig. 5B).

4.2.2 Infer monkeys' mutual predictions in a non-cooperative task

The "chicken" game, a non-cooperative task blending coordination and conflict, offers insights into players' motivations and tendencies [17]. Two monkeys are presented with options: go Straight (S)

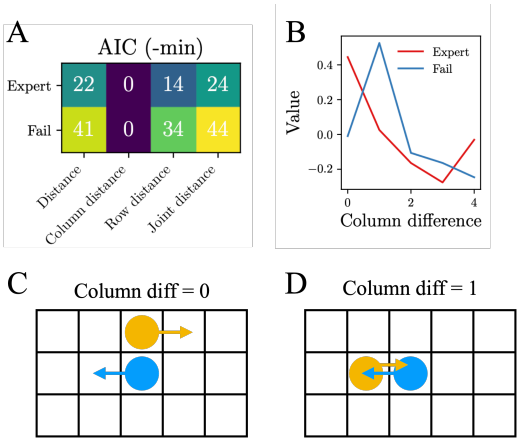


Figure 5: **MAIRL identified the most crucial interaction term related to the task setup**. (A) Model fitness measured by AIC for models with different interaction terms. The interaction map based on column distance performed best.(B) Value attached to the column distance between two agents. Expert trajectories prioritized states where the two agents pass each other (C), while failed trajectories focused on a column distance of one grid apart, suggesting potential collisions and rebounds (D).

or Yield (Y), resulting in four potential outcomes (Fig. 6A and B):(i) S-S leads to a crash with zero reward; (ii)/(iii) S-Y yields a significant reward for the bold monkey and a small one for the yielding monkey; (iv) Y-Y results in cooperation with a moderate reward for both. Fig. 6B illustrates the outcome and payoff matrix from one monkey pair (trial number = 29,147), with decisions exhibiting roughly independent behavior, indicated by a mutual information measure of 0.01 bits.

This non-cooperative task is stateless with four actions in the joint action space. Two separate value functions, V_1 and V_2, need to be parameterized. The probability of agent 1's decision could be written as:

$$P_1(a_1) = \sum_{i \in \{S,Y\}} P_1(a_1|\hat{a}_2 = i) P_1(\hat{a}_2 = i) = \sum_{i \in \{S,Y\}} \pi_1(a_1|a_2 = i) P_1(\hat{a}_2 = i) \tag{9}$$

where $\pi_1(a_1, a_2)$ is a softmax transformation of $V_1(a_1, a_2)$ and denote $P_1(\hat{a}_2 = S) = q_1$. Inference details can be found in theSupplementary. The estimated V_i and q_i values (Fig. 6C) shows that Monkey 1 estimates that its opponent has a lower probability of going straight. Consequently, Monkey 1 is bolder and chooses to go straight more often. Interestingly, the behavior of this monkey matches its higher social status as determined by a separate confrontation experiment.

Notably, the authors of these experiments [17] developed a "hybrid RL and strategic learning model" to determine one animal's prediction about another, which is equivalent to a constrained MARL model with independent control, mutual logistic prediction and hard Q-learning update. Compared to the this model, our MAIRL model are more generally applicable to more complex predictions in social interactions and state-dependent behaviors.

5 Discussion

We used MARL and recursive reasoning to model goal-directed reward representation and levels of cognitive hierarchy levels in social tasks and applied IRL to infer value functions from expert trajectories and determine the level of reasoning by model comparison. We validated our new approach using both simulated tasks and experimental datasets. Unlike previous methods, our approach disentangles agents' perception of task-related goal values from their recursive reasoning about others. This enables us to dissect the distinct computations underlying social cognition and invites future investigation into their underlying neural mechanisms.

This work presents a compelling adaptation of MARL theories to model social interactions, propelling research in neuroscience and cognitive sciences. By integrating advanced machine learning techniques and theoretical guarantees, we provide novel insights into the computations underlying complex social dynamics. Future directions include expanding the algorithm to accommodate continuous state and action spaces and eliminating assumptions of simultaneous moves, as demonstrated in the PR2Q algorithm [26]. Additionally, there is potential for exploring tasks involving equilibrium solution spaces [22][30]. Our approach and its applications hold great promise for advancing our understanding of social behavior and the human mind.

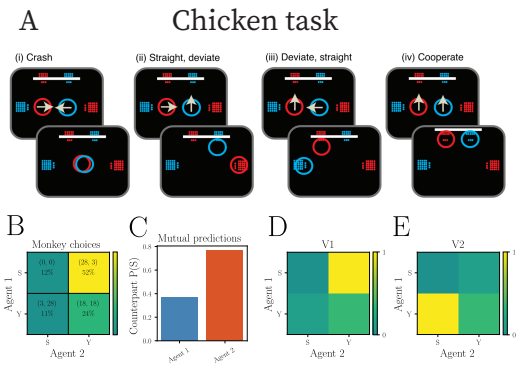


Figure 6: **Monkey's mutual predictions are related to their social hierarchy in a non-cooperative task.** (A) Two monkeys (red and blue circles) choose to go Straight (S) or Yield (Y) in a non-cooperative chicken task. (adapted from [17]) (B) The choices of a pair of monkeys from the experiment (Fig. 2A in [17]). The value pairs in parenthesis indicate the pay-off matrix. (C) (D) (E) Estimated mutual predictions and value maps using a generative model of independent control with prediction. Mutual predictions are the probability of Monkey 2 choosing S as predicted by Monkey 1 and vice versa. Monkey 1 chooses to go straight more often because it is unlikely for Monkey 2 to go straight. This strategy matches their social hierarchy where Monkey 1 has a higher rank than Monkey 2.

Limitations. Our comparison did not appear to distinguish between the two predictive models in independent control (Fig. 3A), possibly due to the simplicity of the task that does not require any sophisticated prediction. Additionally, we opted not to model cognitive processes beyond the first level of reasoning, as they are not required by the tasks of interest. However, more recursive levels of reasoning are feasible and can be used for more complex social scenarios.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

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

[3] S. Arora and P. Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

[4] Z. Ashwood, A. Jha, and J. W. Pillow. Dynamic inverse reinforcement learning for characterizing animal behavior. *Advances in Neural Information Processing Systems*, 35:29663–29676, 2022.

[5] C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3): 329–349, 2009.

[6] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064, 2017.

[7] L. Buşoniu, R. Babuška, and B. De Schutter. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1*, pages 183–221, 2010.

[8] R. Campos-Rodriguez, L. Gonzalez-Jimenez, F. Cervantes-Alvarez, F. Amezcua-Garcia, and M. Fernandez-Garcia. Multiagent systems in automotive applications. *Multi-agent Systems*, page 43, 2017.

[9] H. De Weerd, R. Verbrugge, and B. Verheij. How much does it help to know what she knows you know? an agent-based simulation study. *Artificial Intelligence*, 199:67–92, 2013.

[10] F. Derakhshan and S. Yousefi. A review on the applications of multiagent systems in wireless sensor networks. *International Journal of Distributed Sensor Networks*, 15(5):1550147719850767, 2019.

[11] C. D. Frith and U. Frith. Mechanisms of social cognition. *Annual review of psychology*, 63:287–313, 2012.

[12] U. Frith and C. D. Frith. Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):459–473, 2003.

[13] M. K. Ho, J. MacGlashan, A. Greenwald, M. L. Littman, E. Hilliard, C. Trimbach, S. Brawner, J. Tenenbaum, M. Kleiman-Weiner, and J. L. Austerweil. Feature-based joint planning and norm learning in collaborative games. In *CogSci*, 2016.

[14] J. Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019.

[15] J. M. Keynes. The general theory of employment. *The quarterly journal of economics*, 51(2):209–223, 1937.

[16] K. Khalvati, S. A. Park, J.-C. Dreher, and R. P. Rao. A probabilistic model of social decision making based on reward maximization. *Advances in Neural Information Processing Systems*, 29, 2016.

[17] W. S. Ong, S. Madlon-Kay, and M. L. Platt. Neuronal correlates of strategic cooperation in monkeys. *Nature neuroscience*, 24(1):116–128, 2021.

[18] G. Papoudakis, F. Christianos, L. Schäfer, and S. V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. *arXiv preprint arXiv:2006.07869*, 2020.

[19] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736, 2006.

[20] E. Shakshuki and M. Reid. Multi-agent system applications in healthcare: current technology and future roadmap. *Procedia Computer Science*, 52:252–261, 2015.

[21] S. Shuvaev, E. Amelchenko, D. Smagin, N. Kudryavtseva, G. Enikolopov, and A. Koulakov. A normative theory of social conflict. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] J. Song, H. Ren, D. Sadigh, and S. Ermon. Multi-agent generative adversarial imitation learning. *Advances in neural information processing systems*, 31, 2018.

[23] R. Tejwani, Y.-L. Kuo, T. Shu, B. Katz, and A. Barbu. Social interactions as recursive mdps. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 949–958. PMLR, 08–11 Nov 2022. URL `https://proceedings.mlr.press/v164/tejwani22a.html`.

[24] T. Ullman, C. Baker, O. Macindoe, O. Evans, N. Goodman, and J. Tenenbaum. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22, 2009.

[25] J. Velez-Ginorio, M. H. Siegel, J. B. Tenenbaum, and J. Jara-Ettinger. Interpreting actions by attributing compositional desires. In *CogSci*, 2017.

[26] Y. Wen, Y. Yang, R. Luo, J. Wang, and W. Pan. Probabilistic recursive reasoning for multi-agent reinforcement learning. *arXiv preprint arXiv:1901.09207*, 2019.

[27] M. P. Woodward and R. J. Wood. Learning from humans as an i-pomdp. *arXiv preprint arXiv:1204.0274*, 2012.

[28] Y. Yang. *Many-agent reinforcement learning*. PhD thesis, UCL (University College London), 2021.

[29] W. Yoshida, R. J. Dolan, and K. J. Friston. Game theory of mind. *PLoS computational biology*, 4(12): e1000254, 2008.

[30] L. Yu, J. Song, and S. Ermon. Multi-agent adversarial inverse reinforcement learning. In *International Conference on Machine Learning*, pages 7194–7201. PMLR, 2019.

[31] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[32] B. D. Ziebart, J. A. Bagnell, and A. K. Dey. Modeling interaction via the principle of maximum causal entropy. 2010.

A Supplementary details

A.1 Simulation details of the cooperative foraging task

The arena contains a 5-by-5 grid with deterministic action transitions, resulting in 625 states and 25 actions in the defined two-agent MDP. Two terminal states are defined by the simultaneous presence of both agents at the same reward location. If an agent hits the boundary, only its own state is reset to the previous time step. The reward strength of two locations was sampled from $\mathcal{N}(2, \sigma_0^2)$ with $\sigma_0 = 0.01$. The average reward strength was handpicked to be 2 to enable a balance between sampling from all state-action pairs and emphasizing the importance of centralized control.

At the beginning of each trial, both agents are placed randomly in the arena. Their exploration policy was generated using value iteration and the chosen policy execution model from Table 1. For all datasets, 500 trials were simulated with a maximum length of 200 steps. The discount value γ was set to 0.9.

A.2 Inference details for cooperative foraging

We used 80% trajectories for model fitting and the other 20% to calculate the test log-likelihood as model evidence. The preset hyperparameters include: the future discount factor $\gamma = 0.9$ and the covariance strength of map weights $\sigma_0 = \sigma_1 = 0.01$. In this way, we're imposing a strong prior of the map weights to be around one, for a better comparison of the recovered maps.

Other hyperparameters were searched among the range provided in Table S1 in a centralized dataset inferred using centralized model assumptions. We picked $K = 1, \eta_\alpha = 0.01, \eta_{map} = 0.005, \lambda_1 = 5, \lambda_2 = 1$ based on training stability, model evidence and map interpretability.

K	number of maps	[0,1]
η_α	learning rate for weights	[0.001, 0.005, 0.01]
η_{map}	learning rate for the map	[0.001, 0.005, 0.01]
λ_1	L2-penalty (prior) for individual maps	[0,1,5]
λ_2	L2-penalty (prior) for the interaction map	[0,1]

Table S1: Range of hyperparameters tested

We initialized the weights to be one, the individual maps to be uniformed sampled from 0 to 1, and the location difference map to be $1/(d+1)$ where d is the square Euclidean distance between two agents. We used Adam optimizer with learning rate given above for iterate for 400 epochs. Most loss curves would plateau around 200 epochs. Example loss curves for the training and test set were shown in Fig.S5.

For 500 expert trajectories with approximately 30 decisions per trajectory, the inference process took around 5 minutes on a single CPU thread (Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz).

A.3 Inference details for the hallway task

For the hallway task, most hyperparameters are the same as the cooperative foraging task except that we searched for $\lambda_1 \in [0, 1, 2, 5]$ and $\lambda_2 \in [0, 1, 2]$ using the success trajectories. We used $\lambda_1 = 2$ and $\lambda_2 = 1$.

A.4 Inference details for the chicken task

Unlike the previous two tasks, this non-cooperative task is stateless with four actions in the joint action space. Two separate value functions, V_1 and V_2, need to be parameterized. Value decomposition was not used due to limited number of parameters to estimate. The probability of agent 1's decision could be written as:

$$P_1(a_1) = \sum_{i \in \{S,Y\}} P_1(a_1|\hat{a}_2 = i) P_1(\hat{a}_2 = i) = \sum_{i \in \{S,Y\}} \pi_1(a_1|a_2 = i) P_1(\hat{a}_2 = i) \tag{10}$$

where $\pi_1(a_1, a_2)$ is a softmax transformation of $V_1(a_1, a_2)$ and denote $P_1(\hat{a}_2 = S) = q_1$. Therefore, $P_i(a_i)$ is parameterized by five parameters each while the observed choice probability only offers two observation points. This creates an under-determined system. So we posed some constrains over the value function according to the task setup. We constrained V_1 to be the transpose of V_2 and $V_1(S, Y) > V_1(Y, Y) > V_1(Y, S) > V_1(S, S) = 0$ according to the task's nature. In this way, we could have a faithful estimation of q_1 and q_2, which is usually the biggest ambiguity in these social interactions.

A.5 Validity of value decomposition through numerical optimization

We tested whether the joint value function $r(s_1, s_2) : \mathcal{S} \to \mathbb{R}$ can be reconstructed with individual value maps $m(s_1) : \mathcal{S}_1 \to \mathbb{R}$, $n(s_2) : \mathcal{S}_2 \to \mathbb{R}$ and a value map of distance $\phi(\|s_1 - s_2\|^2)$. In general, we use $O(|\mathcal{S}|)$ to approximate $O(|\mathcal{S}|^2)$, which may not always be possible. But in the case of cooperative foraging, we used an optimzation procedure to show it is possible mostly because the reward function is sparse.

If we denote the two target location as t_1 and t_2, $r = \delta_{(s_1,s_2)\in\{(t_1,t_1),(t_2,t_2)\}}(s_1, s_2)$. We designed a optimization problem as the loss function designed as the reconstruction error:

$$L = \frac{1}{|\mathcal{S}|} \sum_{s_1,s_2} (r - m - n - \phi)^2 + \lambda_1 \|m\|^2 + \lambda_1 \|n\|^2 + \lambda_2 \|\phi\|^2 \tag{11}$$

We dropped the map weight as this reconstruction is only using one set of maps and the weight could be absorbed into the absolute value of map strength. We used an AdAM optimizer with learning rate equals to 0.05 to perform gradient descent. All parameters are initialized at 1. The loss function achieved almost zeros after 20 iterations (Fig.S1 upper left). The recovered maps was also shown in Fig.S1. It is worth noticing that individual maps highlights the reward locations and the magnitude of difference map is much higher than individual maps.

B Supplementary figures

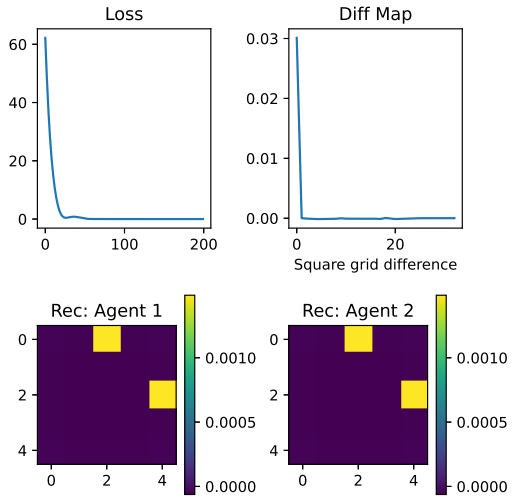


Figure S1: **Value decomposition of a joint map**

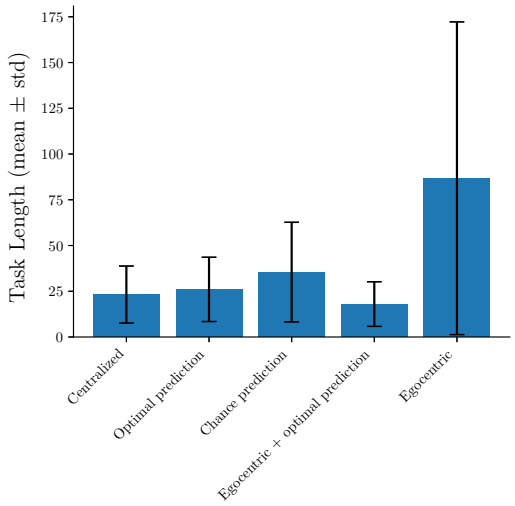


Figure S2: **Task length simulated using different models**

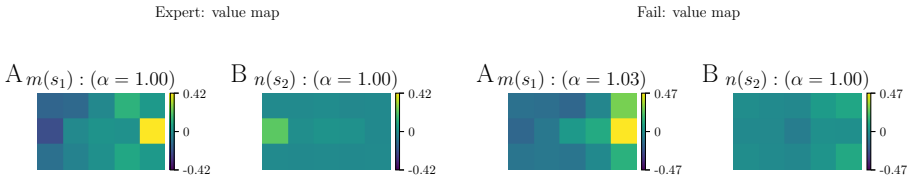


Figure S3: **Expert and failed trajectories fit using independent control with chance prediction**

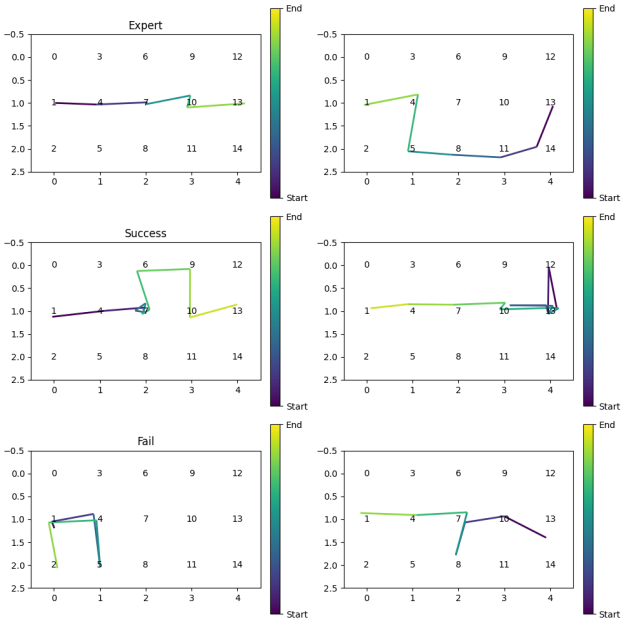


Figure S4: **Example expert (upper), success (middle) and fail (bottom) trajectories in the hallway task**

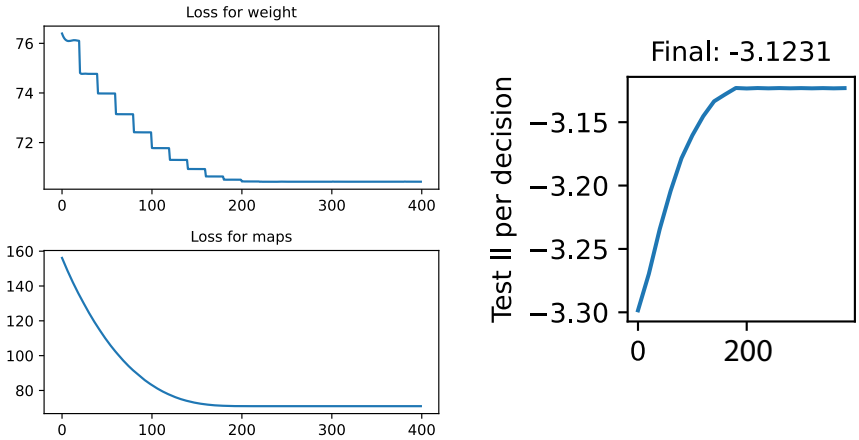


Figure S5: **Example loss curves for a centralized dataset inferred using centralized model assumptions.** Upper left: Loss (LL+prior) curve for weights. The curve is step-wise because we alternated optimizing for weights and maps and there is little room for weight change. Lower left: Loss (LL+L2 penalty) for maps. The curve plateaued around 200 epochs. Right: Test LL per decision along training.