Unveiling value functions in social cognition with multi-agent inverse reinforcement learning

Yusi Chen[1,2,*], Yuan Cheng[3], Myungji Kwak[3], Angela Radulescu[4], Herbert Zheng Wu[3,*]

[1]Computational Neuroscience Center, University of Washington, Seattle, WA, USA
[2]Institute of Neuroscience, Chinese Academy of Sciences, Shanghai, China
[3]Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[4]Department of Psychiatry and Center for Computational Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[*]Correspondence: chenyusi151201@gmail.com (Y.C.), zheng.wu@mssm.edu (H.Z.W.)

Abstract

Social behavior requires individuals to consider not only their own goals but also those of others. Latent value functions that encode such goals can be recovered from behavior using inverse reinforcement learning in single-agent settings. However, extending it to multi-agent interactions is challenging, because value functions are defined over joint state spaces that grow exponentially with the number of agents. Existing approaches often manage this complexity by imposing strong structural assumptions about social interactions, thereby limiting their applicability and interpretability. Here we show that joint value functions governing social interactions can be effectively represented through value decomposition into individual value maps for each agent and low-dimensional interaction terms. We develop a multi-agent inverse reinforcement learning framework (MAIRL) to infer these representations from behavior. In mouse and primate social tasks, MAIRL reveals interpretable value maps that are conditioned on the distinct social roles animals play during group behavior. Together, these results establish MAIRL as an interpretable and scalable framework for identifying latent value representations guiding multi-agent behavior across species.

1 Introduction

Social interactions are central to survival and daily life [1–5], yet deciphering the computations that govern them remains a fundamental challenge. In these settings, one's decisions both influence and depend on the actions of others, making it difficult to determine how the brain evaluates outcomes and integrates social information. Inverse reinforcement learning (IRL) is a powerful approach for studying single-agent behavior by inferring its latent value functions that encode an agent's goals [6–9]. This framework has been instrumental in linking behavior to neural representations [10, 11] and in understanding how these computations break down in neuropsychiatric disorders [12, 13]. Extending IRL to social interactions is therefore essential for uncovering the computational principles underlying social cognition and its dysfunction.

However, applying this framework to social interactions introduces key challenges: the latent value functions are defined over the joint state space of interacting agents, which grows exponentially with the number of agents. To manage this complexity, many existing cognitive and neuroscience models rely on strong structural simplifications, limiting their applicability across diverse social tasks. One common strategy embeds social influences as latent variables within a single-agent Partially Observed Markov Decision Process (POMDP) framework [14–17]. While effective for capturing behavior, these approaches obscure the structure of social interactions by collapsing them into unobserved latent factors. Other approaches [18, 19] adopted multi-agent Markov Decision Processes (MDPs) to infer simplified, often binary, social intentions to improve task performance, but do not recover the full value structure. Related work [20] assumes distinct self-value functions across cognitive hierarchies, conflating task-related value structure with agents' recursive reasoning. These limitations highlight the need for a framework that can recover the complete and interpretable value functions directly from multi-agent behavior. This would enable direct inference of how agents represent both the environment and their social partners, disentangling task structure from social reasoning and allowing quantitative comparison of strategies.

To enable inverse inference of latent value functions from behavior, we first construct a generative forward model that links value functions to observable social decisions under well-defined computational assumptions. We formalize social tasks within a multi-agent reinforcement learning (MARL) framework [21], which specifies how multiple agents act and interact in a shared environment. We show that joint value functions governing social interactions can be efficiently represented through value decomposition, which factorizes them into individual value components for each agent and low-dimensional interaction terms. This representation reduces the dimensionality of the inference problem while preserving task-relevant interactions between agents, enabling tractable and interpretable recovery of joint value functions from behavior. We show analytically and numerically that, in typical social tasks where joint rewards are sparse, this decomposition accurately reconstructs the full joint value function. Within this framework, task structure is encoded in the joint value function, while agents' cognitive hierarchy is captured by recursive reasoning [22], with recursion depth corresponding to theory-of-mind capacity.

Building on this formulation, we develop a multi-agent inverse reinforcement learning (MAIRL) framework [6, 7, 23] that infers decomposed value representations from observed behavior, thereby recovering agents' internal representations of both the environment and their social partners. This inference algorithm accurately recovers the ground-truth value functions in simulated social foraging tasks, validating the approach under controlled conditions. We then apply MAIRL to empirical data from mouse cooperative behavior, revealing how asymmetric value learning across agents gives rise to leadership and followership. In a non-cooperative primate task, MAIRL uncovers mutual prediction signals correlated with social hierarchy. Together, these results demonstrate that value decomposition through MAIRL enables interpretable, scalable inference of the goals and strategies underlying social interactions across tasks and species.

2 Results

2.1 MAIRL with value decomposition

Social interactions between two agents can be modeled as a two-agent Markov Decision Process (MDP)[1] The joint state space was denoted as $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$, with joint state $s = (s_1, s_2)$ combining two agents' individual states, and the joint action space as $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$, with joint action $a = (a_1, a_2)$. The transition kernel $P(s' \mid s, a)$ specifies the environment dynamics, r^i denotes the reward function of agent i, and $\gamma \in [0, 1]$ is the discount factor. The reward function r^i assigns a scalar value to agent i when all agents transition from joint state $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ under joint action $a \in \mathcal{A}$. Consequently, its domain is defined over the joint state–action space, whose cardinality grows exponentially with the number of agents. When reward functions are identical across agents ($r = r^1 = r^2$), the formulation captures cooperative interactions; when they differ, it accommodates competitive or mixed-motive social behaviors. For clarity, we focus on the identical-interest setting and omit agent superscripts below.

In inverse reinforcement learning [6, 7, 24], given a known environment $\{P, \mathcal{S}, \mathcal{A}, \gamma\}$ and expert trajectories $D = \{(s_t, a_t)\}_{t=0}^{T}$, the goal is to recover a reward functions r that maximize the likelihood $P(D|r)$. Under a conditional independence assumption, the trajectory likelihood factorizes as

$$P(D|r) = \prod_{t=0}^{T} \pi_r(a_t|s_t)\, p(s_t)$$

where policy π is induced by the underlying value functions and thus implicitly parameterized by reward r. Details of the problem formulation could be found in Section 4.1.

A key challenge is that joint reward functions in multi-agent settings are defined over a space that grows exponentially with the number of agents, making them difficult to compute and interpret. Here we show that these functions can be effectively represented through **value decomposition**, which expresses them as individual components of each agent and low-dimensional interaction terms:

$$r(s_1, s_2) = \alpha_1 m(s_1) + \alpha_2 n(s_2) + \alpha_3 \phi(d), \tag{1}$$

where m and n denote value maps over individual states, $\phi(d)$ is an interaction term depending on the behavior (e.g., Dijkstra distance d, defined as number of minimal steps), and α_k are scalar weights. This decomposition reduces the parameter space from $|S_1||S_2|$ to $|S_1| + |S_2| + |d|$, while preserving the task-relevant structure of social interactions. Although we refer to this formulation as value decomposition following standard terminology, the inferred quantity in our framework is the reward function; the resulting value function inherits the same structured decomposition through reinforcement learning dynamics. In the following, we use the terms reward and value interchangeably when referring to their shared decomposed structure.

Building on this decomposed value representation, we develop a probabilistic generative model of the joint policy $\pi(a \mid s)$ by integrating a multi-agent reinforcement learning framework with maximum-entropy value iteration [6, 25] (Section 4.3) and coordination strategies (Section 4.4). Using this generative model summarized in Fig. 1A, we can then infer the target reward function that best explains the observed behavior using the proposed inference algorithm (Section 4.5).

2.2 Tractable reconstruction of joint reward functions via value decomposition

To illustrate how value decomposition recovers interpretable latent structure in social tasks, we consider a set of canonical multi-agent foraging paradigms, described in detail below. In each case, the task defines a sparse joint reward function over the combined state space. Although separability of the reward function r is not guaranteed in general, we show (Section 4.2) that for many social tasks the parameters m, n, and ϕ in Eq. 1 can be estimated using an ordinary least squares (OLS) solution (Fig. 1B and Fig. S1). Across

[1] For clarity, we present the two-agent case; the formulation extends naturally to more agents, at the cost of exponential computational growth.

these tasks, we evaluated reconstruction accuracy by measuring the proportion of incorrectly recovered target states as different components of Eq. 1 were included (Fig. 1C). The resulting value maps and interaction terms provide a compact and interpretable description of the latent representations that guide behavior.

In the cooperative match task (Fig. 1D), two agents must reach the same goal location (north or east) simultaneously to obtain reward, resulting in two rewarded joint states. Value decomposition recovers this structure by combining individual value maps that emphasize the goal locations with an interaction term ϕ that peaks at zero distance, reflecting the requirement for spatial proximity. Together, these components reconstruct the sparse joint value function with zero error (Fig. 1C), and provide an interpretable representation of the task: agents value both goal locations and coordination with each other.

In the pursuit task (Fig. 1E), a reward is obtained whenever both agents occupy the same position, independent of their location in the arena. Accordingly, the interaction term alone captures the task structure, encoding value as a function of inter-agent distance, while individual value maps are not required. This demonstrates that value decomposition isolates the relevant latent variable, i.e., agent proximity, underlying the task.

In the cooperative mismatch task (Fig. 1F), agents must reach different goal locations simultaneously to obtain a reward. Here, value decomposition recovers individual value maps that highlight the goal locations, together with an interaction term that encodes the preferred separation between agents. Notably, this interaction term peaks at a distance corresponding to the separation between the two reward zones, reflecting the specific coordination strategy required by the task.

In the three-agent cooperative task (Fig. 1G), reward is delivered when any two agents reach the same goal location simultaneously. Despite the large joint state space ($25^3 = 15{,}625$ states), value decomposition recovers the reward structure using individual value maps and pairwise interaction terms, greatly reducing the parameter space. These components capture both goal-directed behavior and pairwise coordination, demonstrating scalability to interactions with more than two agents.

Together, these examples demonstrate that value decomposition accurately recovers the structure of joint value functions in a compact and interpretable form, revealing how social tasks can be represented in terms of individual goals and interaction constraints. This formulation enables efficient reconstruction with a tractable number of parameters across a range of multi-agent settings.

2.3 Validation of MAIRL in simulated cooperative foraging with heterogeneous strategies

Having established the accuracy of the OLS-based reconstruction, we next asked whether MAIRL can infer decomposed value functions from simulated trajectories of agents with distinct reasoning strategies in the cooperative match task (Fig. 2A). Under independent action selection, the model should recover each agent's value function separately when the agents follow different decision strategies. Following Section 4.4, we simulated two agents with different reasoning policies: an egocentric, theory-of-mind level-0 (ToM-0) agent, which navigates to reward locations without considering its partner's behavior, and a socially aware, theory-of-mind level-1 (ToM-1) agent, which assumes that its partner behaves as a ToM-0 agent and plans accordingly. Example simulated trajectories are shown in Fig. 2B and C. From behavioral observations alone, including trajectory length and port choice, these two strategies are difficult to distinguish.

We then applied MAIRL separately to each agent's trajectories to assess model fit and identify the value components underlying their distinct strategies. First, behavioral fitting using the decomposed value function achieved consistently higher log-likelihood (LL) than the full joint value function across ten random initializations (Fig. 2D). This improvement is likely due to the reduced parameterization of the decomposed model, which makes the optimization less susceptible to local minima. Second, model comparison allowed us to assess which terms in Eq. 5 are necessary for capturing and distinguishing the agents' underlying foraging strategies. The Bayesian Information Criterion (BIC) results of fitting the trajectories from both agents are shown in Fig. 2E and F. A BIC drop greater than 10 is generally considered strong evidence in favor of the lower-BIC model. For the socially aware agent (Fig. 2E), models composed of both self position and partner distance were preferred, whereas for the egocentric

agent, the model containing self position only was sufficient. These selections align with the simulated ground-truth reasoning logic for each agent. The decomposed value map inferred for the socially aware agent features elevated values near the goal locations and in proximal states (Fig. 2G and H), mirroring the numerically decomposed value structure for this task (Fig. 1D) and further validating the inference procedure. Together, these results show that MAIRL recovers agent-specific value structures and supports principled hypothesis testing of behavioral strategies.

2.4 Applications to experimental datasets

After validating our framework in simulations where ground truth is available, we applied the approach to the observed trajectories of two mice performing a cooperative foraging task [26] and tracked how their value representations evolved during learning. This analysis revealed value functions underlying distinct social roles and the unique dynamics in learning. Beyond cooperative settings, we further applied MAIRL to a non-cooperative "chicken" game in monkeys, demonstrating that the method generalizes across a range of social decision-making tasks.

2.4.1 Infer value functions in cooperating mice with distinct social roles

We first applied MAIRL to infer the value representations underlying a cooperative match task performed by a pair of mice (Fig.3A). Similar to the simulated task, the two mice must cooperate by navigating to the same reward zones to obtain rewards [26]. The arena (45cm $\times$ 45cm) was discretized into a 10×10 grid, and animal movements were mapped onto this grid with nine possible actions (stay, up, down, left, right, and four diagonal movements). Example trajectory pairs are shown in Fig. 3B and C, where the animals selected the goal location in the east and coordinated their arrival.

Mice adopt an emergent leader–follower strategy [26] to solve this task: the leader arrives first at the reward zone in most of the trials and has a disproportionate impact on follower's choice. Using MAIRL, we sought to infer the value functions underlying each animal's decisions and to determine how these values relate to their social roles. To this end, we constructed a set of nested models to test whether each component of the decomposed value function was necessary for explaining the behavior. We analyzed data from 27 well-trained animal pairs (Fig.3E-H). For most animals, models incorporating self-location and partner distance provided the best fit, whereas the partner's allocentric location contributed minimally. Notably, partner distance improved model fit more for followers than for leaders. Overall, the best-fitting models significantly outperformed chance-level baselines (Fig. 3F) and accurately predicted final goal choices (Fig. 3G).

We then directly visualized the value maps inferred from observed trajectories during cooperative task. Both leaders and followers (Fig. 4C, middle) assigned higher value to the goal locations. By contrast, followers assigned significantly greater value on proximal partner distance compared to leaders (Fig. 4C, right). Together, these results suggest that animals learn that successful coordination in this task depends primarily on maintaining appropriate mutual distance, with followers relying more heavily on partner distance compared to leaders.

To determine whether MAIRL reliably reveals decision strategies across different behavioral contexts, we compared cooperative foraging to free-run sessions, when animals explore the arena out of the task context while still water-restricted (Fig. 4D). Model comparison (Fig. 4E) indicates that both leaders and followers were best described by models in which decisions depended only on self-location. Although animals continued to assign elevated value to the goal locations (Fig.4F, middle)—likely because these sessions were usually recorded right before task sessions and the well-trained animals continued to visit goal locations in anticipation of potential reward—they no longer incorporated partner distance into their value functions (Fig. 4F, right). This suggests that higher valuation of partner proximity is selective during the cooperative task and absent when animals are not cued to coordinate.

Taken together, these results suggest that successful coordination in this task depends critically on the value representation of partner distance, with followers relying more on partner proximity than leaders. Importantly, these value maps are latent and not directly accessible from behavioral observations

alone, but can be inferred through MAIRL. This demonstrates that the framework enables recovery of interpretable internal representations underlying social decision-making.

2.4.2 Reveal the dynamics of value functions during learning

We next applied MAIRL to track value functions across learning, enabling a closer examination of the dynamics in value assignment. The behavioral trajectories from the same 27 animal pairs were used. Because the value functions were estimated for each training session, we increased the discretization bin size and estimated values on a coarser 5×5 grid.

Across learning, value functions for self-location remained largely stable (Fig. 5A and B), likely reflecting prior training and consistent knowledge of the goal locations. In contrast, the value function for partner distance progressively increased at close proximity over training (Fig. 5C), while the corresponding representation in leaders showed little change (Fig. 5D). These results suggest that followers undergo greater value updating during cooperative learning, selectively reinforcing proximity to their partners.

To quantify this effect at the population level, we pooled the values assigned to proximal partner distance across all 27 animal pairs over learning (Fig. 5E and F). During early learning, followers rapidly increased the value of proximal partner distance, contributing to the increase in coordination success, whereas leaders showed slower and more limited adjustments. Consistent with this pattern, values for proximal partner distance were indistinguishable between followers and leaders at the onset of training (Fig.5G) but became significantly higher in followers after training (Fig. 5H). Together, these results reveal distinct learning strategies for followers and leaders, suggesting that role differentiation emerges through differences in value updating. Importantly, MAIRL reveals these latent value representations that are not directly observable from the behavior, providing a window into the internal computations underlying cooperative learning.

2.4.3 Probing additional behavioral structure via conditioning variables

While the core MAIRL formulation captures social value structure through a compact decomposition, it also enables hypothesis-driven incorporation of additional behavioral details without explicitly expanding the state space. Rather than introducing new state dimensions, we treat candidate variables as conditional modulators of existing value components, allowing us to assess their contribution while preserving tractability. As a proof of concept, we examine whether relative head direction modulates the value associated with inter-animal distance.

Relative head direction (θ) was defined as the angle of the partner from the instantaneous heading of oneself. We model θ not as an independent state variable, but as a conditioning variable that modulates the distance-dependent value term:

$$V(s \mid \theta) = m(s_1) + \phi(d \mid \theta). \tag{2}$$

We discretized the continuous angle θ into three categories—front, left, and right—and fitted these conditioned distance value maps to observed trajectories from the well-trained pairs (Fig. 6). Model comparison revealed that a subset of animals (6 followers and 5 leaders) showed significant improvement after introducing the additional parameters (Fig.3H). For these animals, conditioning on head direction reveals structured modulation of value for partner distance (Fig. 6). Notably, only when the value functions are conditioned on a frontal partner position, followers showed a stronger preference for proximal states than leaders (Fig. 6B), suggesting that followers are more likely to approach and reduce distance when their partner is positioned directly ahead.

This analysis illustrates how MAIRL can flexibly test additional behavioral hypotheses without increasing the dimensionality of the state space. More broadly, it demonstrates that decomposed value representations provide a structured framework for incorporating and evaluating candidate variables, enabling more precise and interpretable characterization of social behavior while preserving computational tractability.

2.4.4 Infer mutual predictions in monkeys during a non-cooperative task

Extending beyond cooperative contexts, we used MAIRL to infer mutual predictions between monkeys in a non-cooperative "chicken" game. This game is a non-cooperative task blending coordination and conflict, offers insights into players' motivations and tendencies [27]. Two monkeys are presented with options: go Straight (S) or Yield (Y), resulting in four potential outcomes (Fig. S2A 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. S2B 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 possible 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:

$$P_1(a_1) = \sum_{i \in \{S,Y\}} P_1(a_1|\hat{a}_2 = i) P_1(\hat{a}_2 = i) = \sum_{i \in \{S,Y\}} \pi_1(a_1|a_2 = i) P_1(\hat{a}_2 = i) \tag{3}$$

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 Section 4.8. The estimated V_i and q_i values (Fig. S2C-E) 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 [27].

Notably, the authors of these experiments [27] 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 this specialized model, our MAIRL model is more generally applicable to more complex predictions and state-dependent behaviors.

3 Discussion

We developed a multi-agent inverse reinforcement learning (MAIRL) framework that enables tractable and interpretable inference of latent value representations in social decision-making. The key advance is a value decomposition formulation that represents joint value functions as structured combinations of individual components and low-dimensional interactions, allowing recovery of high-dimensional social value structure without enumerating the full joint state space. We show theoretically and numerically that this representation accurately reconstructs sparse joint value functions while greatly reducing parameter complexity. Building on this, MAIRL enables direct inference of agents' strategies from behavior: in simulated cooperative tasks, it distinguishes policies with different levels of social reasoning through model comparison. Applying MAIRL to mouse cooperative behavior, we uncover interpretable value maps that reveal social role differentiation and distinct learning dynamics, with followers progressively increasing the value assigned to socially relevant states. Importantly, the interaction terms in the decomposition are user-defined, allowing hypothesis-driven specification of candidate social features and direct testing of their contributions to behavior. MAIRL further supports the incorporation of additional behavioral features, such as head direction, as conditional modulators of value representations without expanding the state space. Extending this framework to a non-cooperative "chicken" game in monkeys, we show that the inferred value functions and mutual predictions capture strategic behavior and relate to social hierarchy, illustrating the generality of the approach across species and social paradigms.

The exponential growth of the joint state space with the number of agents is not only a challenge for computational modeling, but also a fundamental constraint for neural systems operating in social environments. In the social tasks studied here, only a small subset of variables, such as relative spatial position, are behaviorally relevant. In natural social interactions, however, agents must contend with many additional factors, including identity, history, and context, leading to a rapid increase in representational

complexity [28]. Value decomposition provides a principled way to manage this complexity by representing joint value functions in terms of individual components and low-dimensional interaction terms, without enumerating the full joint space. In this view, socially relevant variables act as compact features that capture key aspects of interaction structure. This idea is consistent with the neural representations of partner distance observed during mouse cooperative behavior [26], which resemble the interaction terms recovered by MAIRL. More broadly, such decomposed representations may reflect a general strategy by which the brain achieves dimensionality reduction in multi-agent settings, enabling efficient coordination in complex environments.

The present formulation focuses on a minimal setting that isolates the core computational problem of inferring joint value structure from behavior. We assume full observability of the environment and transition dynamics, allowing us to disentangle value structure from uncertainty about others' states. Extending the framework to partially observable settings would enable modeling how agents infer hidden states and beliefs during interaction. We further consider discrete state and action spaces without imposing a parametric form on the reward function, enabling flexible recovery of value structure. Extensions to continuous domains using appropriate function basis would allow scaling to more complex environments while preserving this interpretability. Finally, the tasks studied here primarily engage low levels of recursive reasoning; applying the framework to settings that require deeper strategic inference would provide a more complete characterization of cognitive hierarchy in social behavior.

Together, these results establish MAIRL as a general framework for uncovering the computations underlying social interactions in biological and artificial systems. The framework naturally extends along several directions enabled by its decomposed structure. First, additional behavioral variables, such as motivation [29], movement kinetics [30], or vocal communication [31], can be incorporated to allow for systematic evaluation of their contribution to social behavior. Second, value decomposition enables scaling to larger groups by reducing parameter growth from exponential to approximately linear in the number of agents, making it possible to study structured interaction graphs and collective dynamics in larger social systems, such as schooling fish or social insects [5, 32, 33]. Third, applying MAIRL to tasks that require deeper strategic reasoning, such as the stag-hunt game [34], would enable quantitative inference of theory-of-mind capacities. These directions highlight that MAIRL provides a flexible and extensible foundation for investigating the computations governing complex social behavior.

Lastly, with the rapid development of large language model (LLM) agents, the proposed framework may also provide a useful perspective for studying interactions in artificial multi-agent systems. Recent advances have enabled multiple LLM-based agents to communicate, negotiate, and coordinate in shared environments, giving rise to increasingly complex collective behaviors [35]. However, the mechanisms governing these interactions often remain difficult to interpret, as policies and internal representations are distributed across high-dimensional neural networks. By explicitly modeling value functions, beliefs, and recursive reasoning, the MAIRL framework offers a principled way to infer latent objectives and predictive structures from observed social behaviors. This approach could help reveal how cooperative or competitive strategies emerge, how social roles differentiate within groups of agents, and how stable social conventions or equilibria arise over time. More broadly, applying such interpretable modeling tools to artificial agent societies may provide a quantitative bridge between studies of biological social cognition and the growing field of multi-agent artificial intelligence, enabling comparative analyses of learning dynamics, coordination strategies, and the emergence of collective behavior across natural and artificial systems.

4 Methods

4.1 Formulation of multi-agent inverse reinforcement learning (MAIRL)

Let's consider a Markov Decision Process [2] 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
- $P = P(s'|s, a) : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ describes the state transition dynamics of the environment, where Δ is a probability measure over $\mathcal{S}$.
- $r^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function that returns a scalar value to the i-th agent for a transition from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$, taking joint action $a \in \mathcal{A}$.
- $\gamma \in [0, 1]$ is the discount factor for future steps.

In inverse reinforcement learning (IRL) [24][6][7], 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 such 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. 4).

$$P(\zeta_i|r) = \prod_{t=0}^{T} \pi(a_t|s_t)p(s_t) \tag{4}$$

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 Fig. 1A.

4.2 Value decomposition

Inferring the reward function $r : \mathcal{S} \to \mathbb{R}$ without assuming a parametric form requires estimating $|\mathcal{S}|$ parameters. In multi-agent settings, this becomes intractable, as the cardinality of $\mathcal{S}$ grows exponentially with the number of agents. To address the issue, noting that r is usually sparse (see examples in Fig. 1D-G), we decompose the joint value map into marginal maps and interactions maps as Eq. 5 where $s_i \in \mathcal{S}_i$ and d stands for a given interaction function between s_1 and s_2 (e.g., Dijkstra distance between agent 1's location s_1 and agent 2's location s_2 in a grid environment).

$$r(s_1, s_2) = \alpha_1 m(s_1) + \alpha_2 n(s_2) + \alpha_3 \phi(d) \tag{5}$$

As a proof of principle, let's first consider the case where the joint value function $r(s_1, s_2) : \mathcal{S} \to \mathbb{R}$ can be approximated by decomposing it into individual value maps $m(s_1) : \mathcal{S}_1 \to \mathbb{R}$, $n(s_2) : \mathcal{S}_2 \to \mathbb{R}$ and an interaction term $\phi(d)$ as a function of Dijkstra distance $d(s_1, s_2)$. While this separability property is not always guaranteed, reconstruction error $r - m - n - \phi$ can be minimized through ordinary least square solution (OLS) of an overdetermined linear system.

Formally, we approximate the joint reward function $r(s_1, s_2)$ by a separable decomposition of the form $m(s_1) + n(s_2) + \phi(d)$. This poses $O(|\mathcal{S}|^2)$ constrains to a system with $O(2 \times |\mathcal{S}|)$ parameters, constructing an overdetermined linear system. As depicted in Fig. S1, the reconstruction is expressed (Eq. 1) as $y = X\beta$ where y is a flattened version of $r(s_1, s_2)$, β is a concatenated vector of parameters in the m, n, ϕ and X is a fixed design matrix depending on the definition of the interaction term d and the

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

grid environment. Since the number of constraints is much larger than the number of parameters, this is an overdetermined system. The ordinary least square (OLS) solution is expressed as $\hat{\beta} = (X^T X)^{-1} X^T y$ and the error vector evaluated as $(I - X(X^T X)^{-1} X^T)y$, denoted as My. The reconstruction error is the projection of vector y onto the null space of M. Since M and y both depend on the task setup and the choice of the interaction term, we may calculate the reconstruction error before performing the inference (Fig. 1C-G).

4.3 Policy parameterization

After parameterizing the reward function, we adopted a differentiable maximum entropy policy [6][25] to obtain a probabilistic mapping from reward functions to policies:

$$\pi(a|s) = \frac{\exp Q(s,a)}{\sum_{a' \in \mathcal{A}} \exp Q(s,a')} \tag{6}$$

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

$$Q(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) \log(\sum_{a' \in \mathcal{A}} \exp Q(s',a')) \tag{7}$$

Note that a temperature term is 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 from observations.

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. In most realistic scenarios, however, agents have their own objectives and are not governed by a centralized controller. We therefore introduce separate reward functions, value functions, and policies r^i, Q^i and $\pi^i(a^i|s)$ for each agent. If we assume that agents act independently and simultaneously, their joint policy could be expressed as $\Pi_i \pi^i(a^i|s)$. This will be the primary policy formulation to describe multiple agents in this work.

4.4 Coordination and theory-of-mind (ToM) levels

Another source of complexity in modeling multi-agent interactions arises from the need to account for coordination and recursive reasoning between agents. Each agent's behavior depends not only on the environment but also on its beliefs about the strategies of others, introducing multiple levels of reasoning that must be modeled. For clarity, we focus on cooperative tasks in which agents share an identical value function. We further assume that each agent's state is fully observable to the others. The theory-of-mind (ToM) levels considered in this work, corresponding to different depths of recursive reasoning, are summarized in Table 1.

In the fully coordinated scenario, agents behave as if controlled by a centralized planner, executing the optimal **joint action** defined by Eq. 6. In the following, π denotes this optimal joint policy, which may differ from the policies actually implemented by individual agents.

In a less coordinated scenario, agents independently execute their policy, which is often the case in many simultaneous action games. Mathematically, we have conditional independence of individual policies $P(a|s) = P(a_1|s)P(a_2|s)$. The question then becomes how each agent chooses its own action based solely on state information, without knowing the other agents' action plans. The answer can be formulated through a probabilistic view of recursive reasoning, as described below:

For a ToM level-0 agent, which 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 can be generated as soft value iteration of its own map function. For a level-1 or higher agent, 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 express the policy of agent 1 as:

$$P(a_1|s) = \sum_i P(a_1, \hat{a}_2 = i|s) = \sum_i P(a_1|\hat{a}_2 = i, s) P(\hat{a}_2 = i|s) \tag{8}$$

Table 1: Generative models with different coordination levels.

Coordination level	Mutual predictions	Agent i's action probability
Independent control	egocentric	$P(a_i \mid s_i)$
Independent control	chance prediction	$\lvert\mathcal{A}_{-i}\rvert^{-1} \sum_{a_{-i}} \pi(a_i \mid a_{-i}, s)$
Independent control	optimal prediction	$\sum_{a_{-i}} \pi(a_i, a_{-i} \mid s)$

Since a level-1 agent acts optimally for the joint task based on its own predictions, the first term is essentially the conditional optimal policy $P(a_1|\hat{a}_2, s) = \pi(a_1|a_2, s)$. Thus, an agent's policy is determined jointly by its prediction about its counterpart's behavior and the underlying joint value function. The recursive level of this prediction defines the agent's theory-of-mind (ToM) level. For example, if agent 1 assumes agent 2 to be a level-0 agent (regardless of whether agent 2 is or not), then agent 1 has a ToM 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 ToM level of two. Even higher levels of reasoning can be constructed using this recursive process. However, prior work suggests that human and animal behavior in strategic settings is typically well described by low levels of recursive reasoning [36–38]. Motivated by this and to maintain model tractability, we focus on low ToM levels that capture basic forms of social inference. Specifically, we consider two limiting parameterizations of mutual prediction under independent control: one in which agents make no informed predictions about their partners, and one in which agents assume their partners act according to the optimal marginal policy. These cases provide lower and upper bounds on coordination under this framework:

- If agent 1 has no information about agent 2's policy, it 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, it also acts on the marginal probability of the optimal joint policy. (i.e. $P(\hat{a}_2|s) = \pi(a_2|s)$), then $P(a_1|s) = \sum_i P(a_1|a_2 = i, s)\pi(a_2 = i|s) = \pi(a_1|s)$. Intuitively, when an agent correctly predicts that its partner follows the optimal policy, integrating over these predictions recovers its own marginal behavior under the optimal joint policy.

Most results in this paper, except Fig. S2, were inferred under the assumption of independent control with optimal prediction (last row in Table 1). These alternative assumptions about coordination and prediction can be evaluated through model comparison, as described later.

4.5 Inference procedure

After establishing a generative model based on the chosen hypothesis, our objective is to learn the map weights α_k , individual value maps m, n and the interactions map ϕ that maximize the posterior of the observed expert trajectories $D = \{(s_t, a_t)\}$.

The inference procedure is adapted from Ashwood et al. [7], noting that we have different components for the joint value function. 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 $\text{diag}(\Sigma) = (\sigma_0^2, \sigma_0^2, \sigma_1^2)$, $m, n \sim \mathcal{N}(\mathbf{0}, \Sigma_m)$, $\phi \sim \mathcal{N}(\mathbf{0}, \Sigma_\phi)$. 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 coefficients λ_1 and λ_2 to the map entries. The parameter optimization can be written as:

$$
\begin{aligned}
\boldsymbol{\alpha}^* &= \text{argmax}_\alpha \sum_{(s_t,a_t)\in D} \log P(s_t, a_t) - \frac{1}{2}((\boldsymbol{\alpha} - \mathbf{1})^T \Sigma^{-1} (\boldsymbol{\alpha} - \mathbf{1})) \\
(m^*, n^*, \phi^*) &= \text{argmax}_{m,n,\phi} \sum_{(s_t,a_t)\in D} \log P(s_t, a_t) - \lambda_1 \|m\|^2 - \lambda_1 \|n\|^2 - \lambda_2 \|\phi\|^2
\end{aligned}
\tag{9}
$$

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

Algorithm 1 MAIRL with value decomposition

Input1: MDP information: $\langle \mathcal{S}, \mathcal{A}, P, \gamma \rangle$;
Input2: Expert trajectories D;
Input3: Hyperparameters: σ_0, σ_1, λ_1, λ_2
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. 5);
  Get *optimal joint policy* π using softmax iteration until convergence (Eq. 6 and 7);
  Get *joint policy* $P(a|s)$ based on independent policy execution;
  Update $\boldsymbol{\alpha}$ given reward map fixed (Eq.9, line 1);
  Calculate joint reward based on m, n, ϕ and generate new policy as before;
  Update m, n, ϕ given weights fixed (Eq.9, line 2);
end for
Output: $\boldsymbol{\alpha}, m, n, \phi$

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 2 in a centralized dataset inferred using centralized model assumptions. We picked $\eta_\alpha = 0.01, \eta_{map} = 0.005, \lambda_1 = 5, \lambda_2 = 1$ based on training stability, model evidence and map interpretability.

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 2: Range of hyperparameters tested

We initialized the weights to be one, the individual maps to be uniformly sampled from 0 to 1, and the location difference map to be $1/(d+1)$ where d is the square Dijkstra distance between two agents. We used Adam optimizer with learning rate given above to iterate for 400 epochs. Most loss curves would plateau around 200 epochs. 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).

4.6 Simulation of the cooperative foraging task

For the simulated trajectories shown in Fig. 2, the arena was discretized into a 5-by-5 grid with deterministic action transitions, resulting in 625 states and 25 actions in the two-agent MDP. Two terminal states are defined by the simultaneous presence of both agents at the same reward location either in the north or east. If an agent reaches the boundary of the arena, its position is reset to that of the previous time step. The reward strength at both goal locations was set to 2, as absolute reward magnitude does not affect the inferred value structure. At the beginning of each trial, both agents are placed randomly in the arena. Agents are assumed to act under independent control. Agent 2 follows an egocentric (ToM level-0) policy, whereas agent 1 is modeled as a ToM level-1 agent that predicts agent 2's behavior. In all simulations, 500 trials were generated, each with a maximum length of 200 steps. The discount factor γ was set to 0.9.

4.7 Pre-processing of mouse foraging trajectories

For inference using observed mouse foraging trajectories, each mouse was represented as a massless point anchored at its neck location. The arena was divided into a 10×10 grid with 5cm spatial bins (10cm bins in Fig. 5 and 6). An animal's grid position was determined by assigning its actual location to the nearest grid point. To minimize artifacts introduced by discretizing the arena, we excluded time steps where both animals remained stationary. Each animal had nine possible actions: stay, up, down, left, right, up-left, up-right, down-left, and down-right. The objective was to estimate a value map that, given both animals' current location, maximizes the probability of the observed action. The partner angle in Fig. 6 was defined as the angle of partner's position relative to the agent's own head direction. This angle was discretized into three bins: front (-60 to 60 degree), left (60 to 180 degree) and right (180 to 300 degree), such that $\theta \in \{0, 1, 2\}$. The discretized angle was included as an input to the model. To future limit model complexity, the decomposed joint value function incorporates angle only through the interaction term.

4.8 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.

5 Figures

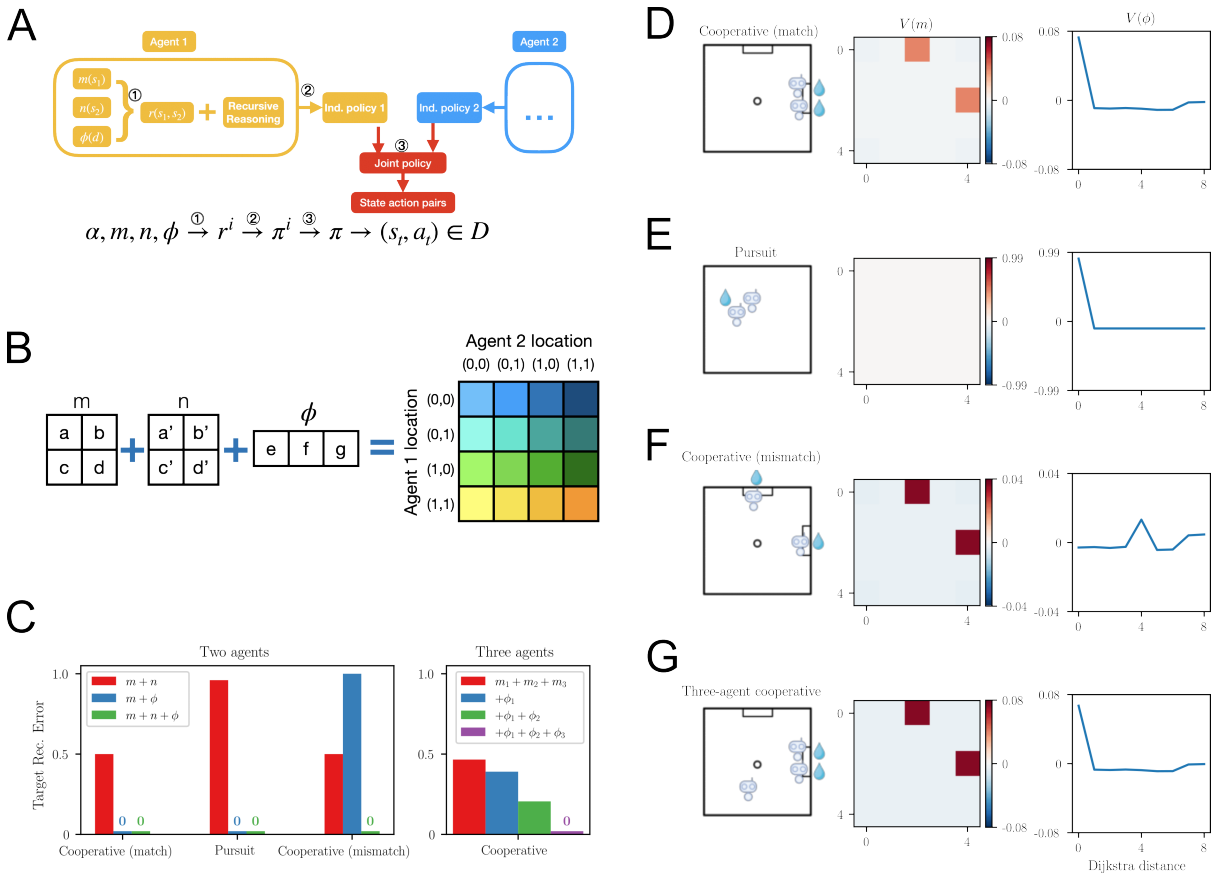


Figure 1: **MAIRL and value decomposition for multiple social tasks.** (A) Model overview. From observed joint state–action pairs (s_t, a_t), we infer each agent's decomposed value functions m, n, ϕ based on a generative model of social decision making. The computation graph highlights the following stages: ① value decomposition (Eq. 1), ② policy generation with soft Q-learning (Section 4.3) and ③ coordination and cognitive levels (Section 4.4). (B) Value decomposition formulated as an overdetermined linear system. The actual reward function of joint locations $r(s_1, s_2)$ is decomposed into summations of functions of each agent's location $m(s_1)$ and $n(s_2)$ and their mutual distance $\phi(d)$. For two agents interact in a 2×2 grid world, 11 parameters $a - g$ from three functions m, n, ϕ (left) are used to reconstruct the joint reward function r on the right with 16 free parameters. The design matrix is defined by the spatial relationships and grid layout (Fig. S1). (C) Target reconstruction error of the joint reward function from the ordinary least squares (OLS) solution of the value decomposition equation. Left, in all two-agent tasks under consideration, $m + n + \phi$ (green bars) suffices to achieve accurate reconstruction of target locations in the joint value function. Right, in a three-agent cooperative task, incorporating each individual's allocentric locations and their pairwise distances allows errorless target location reconstruction (purple bar). (D-G) OLS solutions of decomposed value functions m_i, ϕ_i for different social tasks. Left, schematics for cooperative tasks in two-agent or three-agent cases; Middle: value function of individual's self allocentric location; Right: value function of mutual distance.

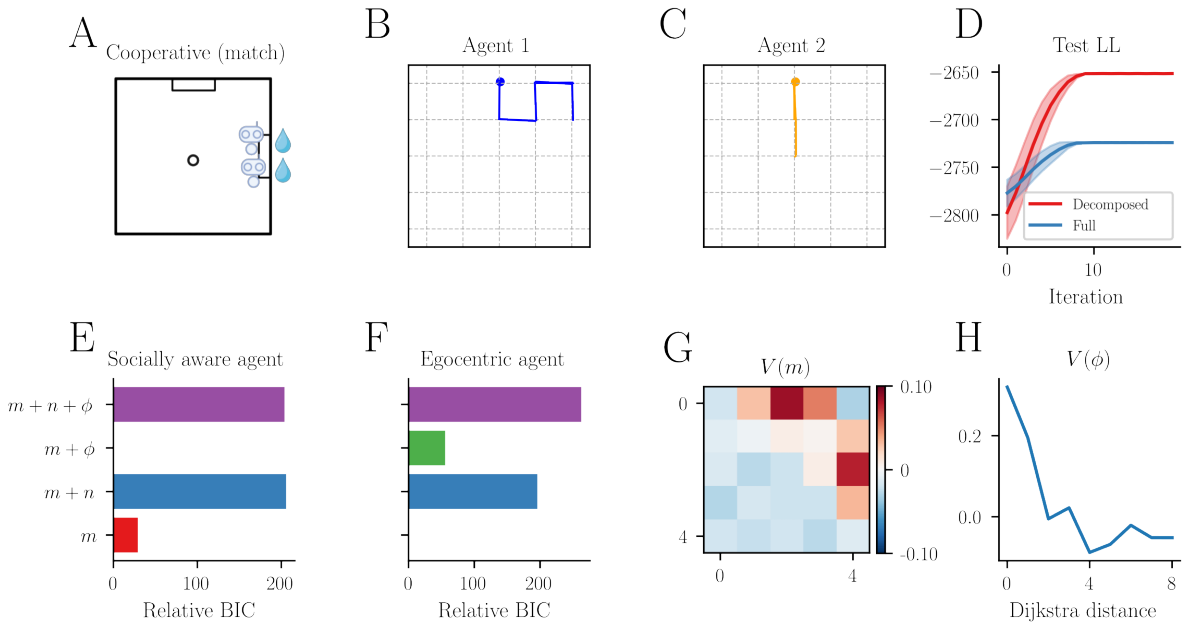
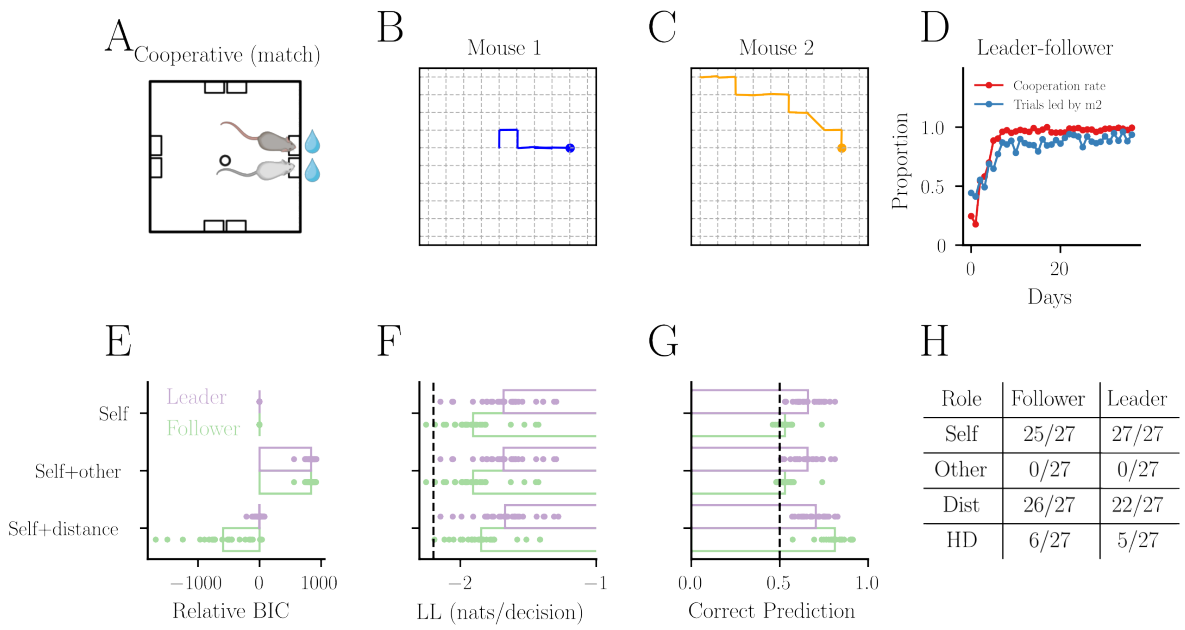


Figure 2: **Recovery of value functions in simulated two-agent cooperative match foraging.** (A) Schematic of a simulated cooperative (match) foraging experiment in which reward is delivered only when both agents reach the target location simultaneously. (B–C) Example simulated trajectories of two agents in a 5×5 grid environment; solid dots indicate trajectory endpoints. (D) Test log-likelihood (LL) of the inference algorithm using either decomposed value parameterization or the full joint value function. Shaded regions denote variability across ten random initializations. The decomposed parameterization achieved higher LL, reflecting improved optimization due to reduced parameter dimensionality. (E-F) Relative Bayesian Information Criterion (BIC) of models with different structural components, fit to trajectories generated by (E) a considerate agent using a ToM-1 policy and (F) an egocentric solo-foraging agent. Each row corresponds to a model in which components of the value decomposition (Eq. 1) are added sequentially, enabling systematic hypothesis testing of behavioral strategies. Relative BIC values are normalized by subtracting the minimum BIC across model hypotheses. (G–H) Inferred value maps for the ToM-1 agent, shown as functions of self-location and mutual distance, revealing strong weighting of target locations and proximal interaction states.



Role	Follower	Leader
Self	25/27	27/27
Other	0/27	0/27
Dist	26/27	22/27
HD	6/27	5/27

Figure 3: **Hypothesis testing of decomposed value models reveals structured interactions and role-dependent strategies in mouse pairs** (A) Schematic of the cooperative foraging task in mice. Target locations are on the east and north. (B-C) Recorded trajectories from animal pairs navigating a 10×10 grid, reconstructed from empirical foraging data. (D) Proportion of cooperative trials and fraction of trials led by a single animal across training days. Over time, animals developed a consistent leader–follower strategy. (E) Relative BIC for 27 animal pairs, normalized to the minimal model containing only self-position value. Purple and green denote model fits for leader and follower animals, respectively. Each row corresponds to a fitting model in which different components in Eq. 1 are added sequentially. Adding allocentric other's position did not improve model fit for either role, whereas including mutual distance significantly improved performance for both. (F) Same format as (E) but for Log-likelihood per decision. The vertical dashed line indicates chance-level. (G) Prediction accuracy of the reward zone selected by animals, based on model-simulated trajectories. The dashed line indicates chance-level accuracy. (H) Number of animals (out of total) showing a significant improvement in model fit as each component was sequentially added. The self position model was first compared against a constant (uniform action) model; additional components were tested against the best-fitting simpler model using chi-square test. Models involving head-direction (HD) is described in Fig. 6.

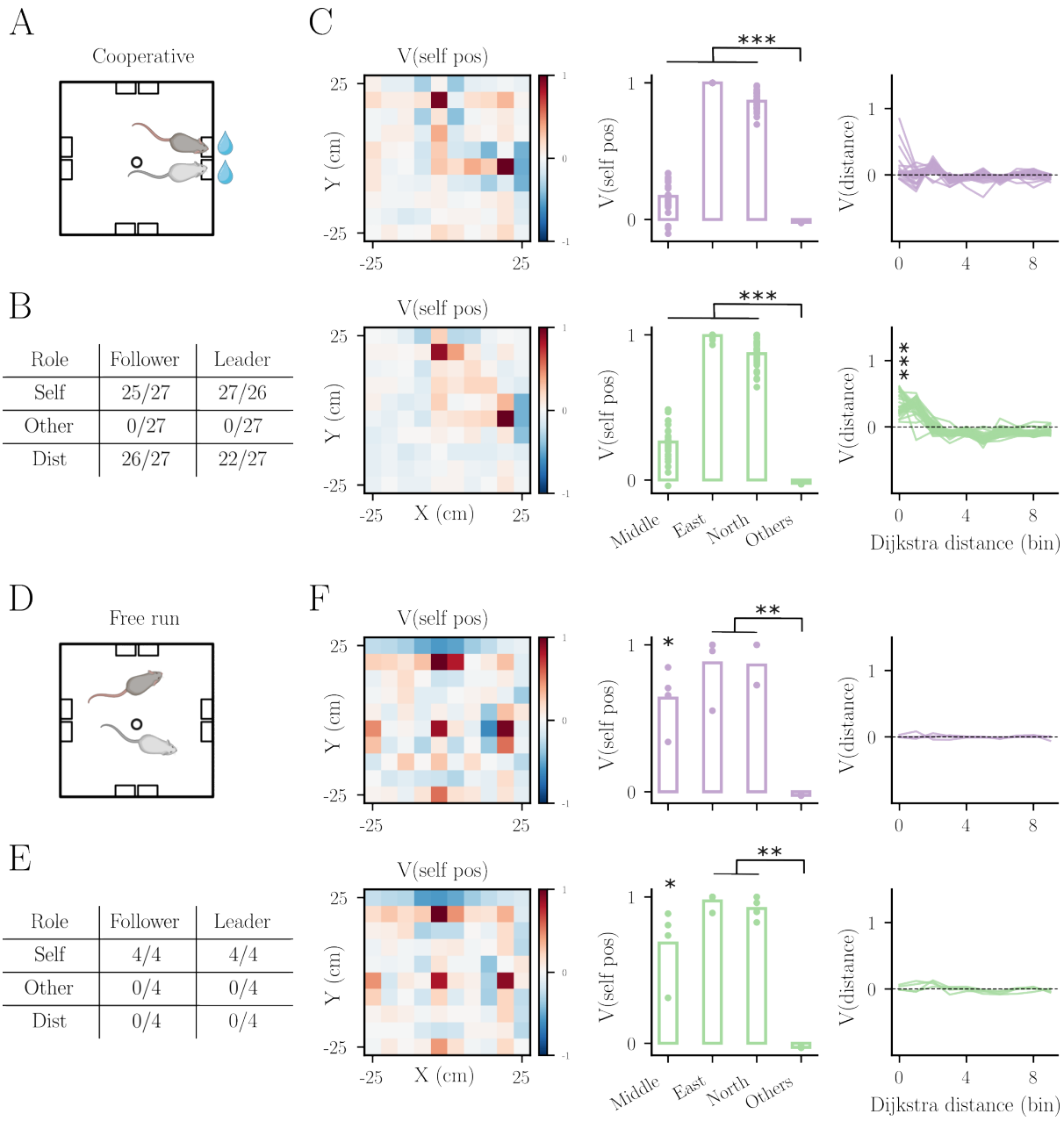


Figure 4: **Value maps inferred from observed trajectories during cooperative foraging and free-run sessions.** *Cooperative task:* (A) Schematic of the cooperative task. (B) Number of animals showing significant improvement in model fit as components were sequentially added. (C) Inferred value functions in correct trials for leaders (top row) and followers (bottom row). Left, example self-position value map in allocentric space from one animal. Middle, value assigned to the middle initiation spot and to the east and north reward ports, and the mean value of all other locations across animals. Both leaders and followers assign significantly higher value to initiation and reward locations (paired t-test, $***P < 10^{-7}, n = 27$). Right, inferred value as a function of inter-agent distance. Followers assign significantly higher value to proximal partner locations than leaders (t-test at distance $= 0$, $P = 6 \times 10^{-13}, n = 27$). *Free-run:* (D–E) Same format as (A–B), but for free-run sessions. (F) Left, example self-position value map. Middle, value assigned to different locations across animals: middle locations show slightly elevated values compared to other positions ($*P < 0.05$), while target locations exhibit significantly higher values ($**P < 0.01, n = 4$). Right, inferred value as a function of inter-agent distance is not significantly different from zero, consistent with the lack of improvement from including the distance component in Fig. 4E.

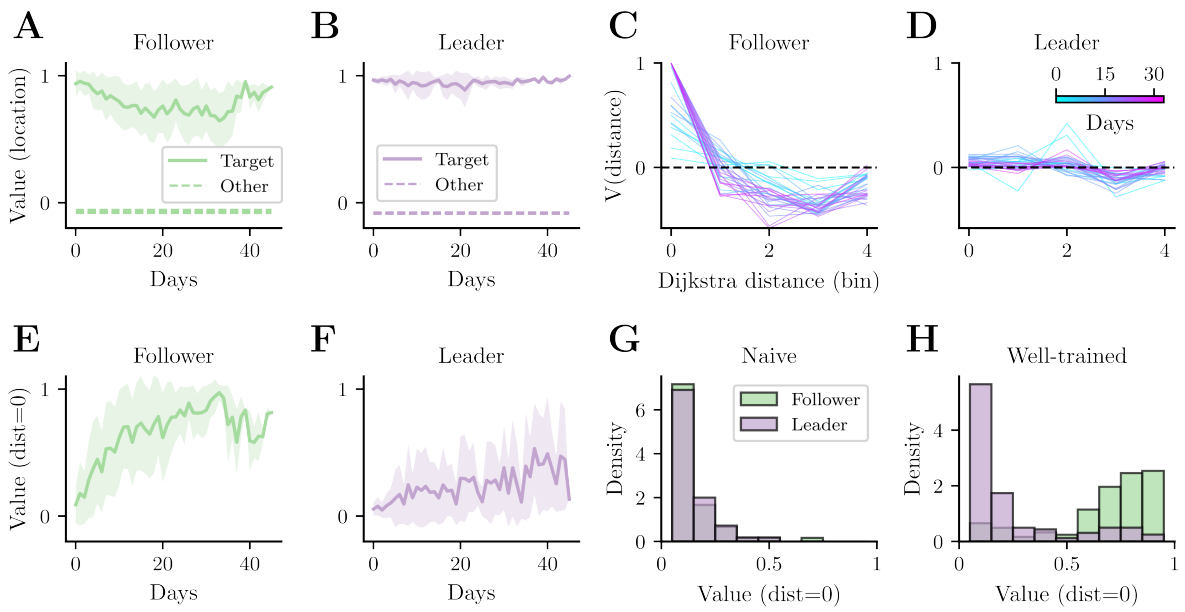


Figure 5: **Dynamics of inferred value functions during learning.** (A–B) Mean inferred value assigned to target locations (north and east) across training days for followers (A) and leaders (B). Solid lines indicate mean values across 27 animals, shaded regions denote ± 1 standard deviation, and dashed lines show the mean value assigned to non-target locations. (C-D) Example inferred value functions as a function of inter-agent distance across learning days for a follower (C) and a leader (D). The follower progressively assigns higher value to proximal states, whereas the leader shows minimal change. (E-F) Value assigned to the proximal state (distance $= 0$) over training days for followers (E) and leaders (F) ($n = 27$). (G-H) Distribution of proximal-state values at early training stage (cooperation rate < 0.6) and at the well-trained stage (cooperation rate > 0.8). Followers exhibited significantly higher proximal-state values than leaders (t-test, $p < 10^{-177}$).

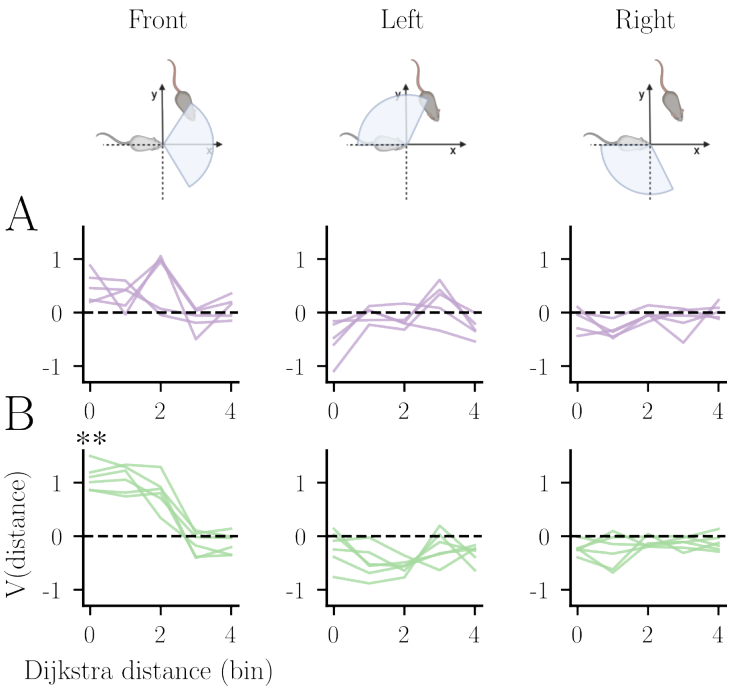


Figure 6: **Distance based value functions conditioned on egocentric head directions.** (A)(B) Leaders' (followers') distance-based value functions conditioned on the partner being in front, to the left, or to the right. A significant role difference emerged only in the front-view condition, where followers assigned greater value to the proximal state (distance $= 0$) than leaders. $^{**}P < 0.01$

References

[1] Robin I. M. Dunbar. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5):178–190, 1998.

[2] Joan B. Silk. Social components of fitness in primate groups. *Science*, 317(5843):1347–1351, 2007.

[3] Ralph Adolphs. The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, 60:693–716, 2009.

[4] Dong V. Wei, Vikram Talwar, and Dayu Lin. Neural circuits of social behaviors: Innate yet flexible. *Neuron*, 109(10):1600–1620, 2021.

[5] F. Engert. Social behavior: A tiny fish with prodigious skills. *Current Biology*, 2025. Study on Danionella cerebrum social interactions.

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

[7] Zoe Ashwood, Aditi Jha, and Jonathan W Pillow. Dynamic inverse reinforcement learning for characterizing animal behavior. *Advances in Neural Information Processing Systems*, 35:29663–29676, 2022.

[8] Feiyang Wu, Ye Zhao, and Anqi Wu. Distributional inverse reinforcement learning. *arXiv preprint arXiv:2510.03013*, 2025.

[9] Jingyang Ke, Feiyang Wu, Jiyi Wang, Jeffrey Markowitz, and Anqi Wu. Inverse reinforcement learning with switching rewards and history dependency for characterizing animal behaviors. *arXiv preprint arXiv:2501.12633*, 2025.

[10] Ethan S. Bromberg-Martin, Yang-Yang Feng, Takaya Ogasawara, J. Kael White, Kaining Zhang, and Ilya E. Monosov. A neural mechanism for conserved value computations integrating information and rewards. *Nature Neuroscience*, 27(1):159–175, January 2024.

[11] Nir Moneta, Shany Grossman, and Nicolas W. Schuck. Representational spaces in orbitofrontal and ventromedial prefrontal cortex: task states, values, and beyond. *Trends in Neurosciences*, 47(12):1055–1069, December 2024.

[12] Quentin J. M. Huys, Tiago V. Maia, and Michael J. Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19:404–413, 2016.

[13] P. Read Montague, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16:72–80, 2012.

[14] Koosha Khalvati, Seongmin A Park, Jean-Claude Dreher, and Rajesh P Rao. A probabilistic model of social decision making based on reward maximization. *Advances in Neural Information Processing Systems*, 29, 2016.

[15] Sergey Shuvaev, Evgeny Amelchenko, Dmitry Smagin, Natalia Kudryavtseva, Grigori Enikolopov, and Alex Koulakov. A normative theory of social conflict. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

[17] Joey Velez-Ginorio, Max H Siegel, Joshua B Tenenbaum, and Julian Jara-Ettinger. Interpreting actions by attributing compositional desires. In *CogSci*, 2017.

[18] Sarah A Wu, Rose E Wang, James A Evans, Joshua B Tenenbaum, David C Parkes, and Max Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021.

[19] Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22, 2009.

[20] Wako Yoshida, Ray J Dolan, and Karl J Friston. Game theory of mind. *PLoS computational biology*, 4(12):e1000254, 2008.

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

[22] Ying Wen, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. Probabilistic recursive reasoning for multi-agent reinforcement learning. *arXiv preprint arXiv:1901.09207*, 2019.

[23] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

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

[25] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. 2010.

[26] Yuan Cheng, Yusi Chen, Myungji Kwak, Ross P. Kempner, Rudramani Singha, Jared Winslow, Runqi Liu, Umais Khan, Tessa Spangler, Alvi Khan, Talmo Pereira, Matthew Whiteway, Evan S. Schaffer, Nuttida Rungratsameetaweemana, Nan Yang, and Herbert Zheng Wu. Asymmetric social representations in the prefrontal cortex for cooperative behavior. *bioRxiv*, 2025.

[27] Wei Song Ong, Seth Madlon-Kay, and Michael L Platt. Neuronal correlates of strategic cooperation in monkeys. *Nature neuroscience*, 24(1):116–128, 2021.

[28] Angela Radulescu, Yeon Soon Shin, and Yael Niv. Human representation learning. *Annual Review of Neuroscience*, 44:253–273, 2021.

[29] W. E. Allen, M. Z. Chen, N. Pichamoorthy, R. H. Tien, M. Pachitariu, L. Luo, and K. Deisseroth. Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science*, 364(6437):eaav3932, 2019.

[30] Yaning Han, Ke Chen, Yunke Wang, Wenhao Liu, Zhouwei Wang, Xiaojing Wang, Chuanliang Han, Jiahui Liao, Kang Huang, Shengyuan Cai, Yiting Huang, Nan Wang, Jinxiu Li, Yangwangzi Song, Jing Li, Guo-Dong Wang, Liping Wang, Yaping Zhang, and Pengfei Wei. Multi-animal 3d social pose estimation, identification and behaviour embedding with a few-shot learning framework. *Nature Machine Intelligence*, 2024.

[31] M. R. Warren, R. S. Clein, M. S. Spurrier, E. D. Roth, and J. P. Neunuebel. A neural basis for ultrasonic vocal communication in mice. *Nature Neuroscience*, 24:1314–1324, 2021.

[32] Iain D. Couzin, Jens Krause, Nigel R. Franks, and Simon A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433:513–516, 2005.

[33] Jo-Hsien Yu, Julia L. Napoli, and Matthew Lovett-Barron. Understanding collective behavior through neurobiology. *Current Opinion in Neurobiology*, 2024.

[34] Sarah F. Brosnan, Brent J. Wilson, and Michael J. Beran. Chimpanzees coordinate in a stag hunt game. *Animal Behaviour*, 82:1–9, 2011.

[35] Mrinmaya Sachan et al. Emergence of sustainable cooperation in a society of llm agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[36] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

[37] Terence C Burnham, David Cesarini, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace. Higher cognitive ability is associated with lower entries in a p-beauty contest. *Journal of Economic Behavior & Organization*, 72(1):171–175, 2009.

[38] Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326, 1995.

A Supplementary figures

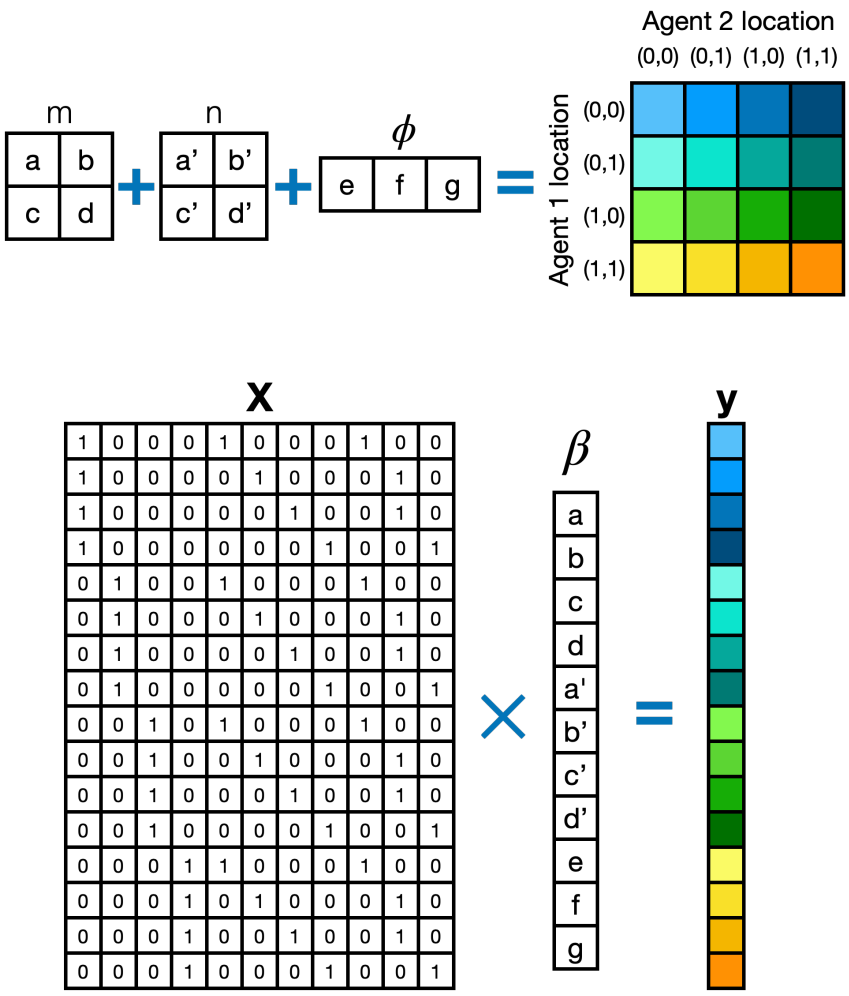


Figure S1: **Value decomposition as an overdetermined linear system.** Upper: for illustration, this task considers two agents interacting in a 2×2 square arena, with the interaction value function ϕ defined over Dijkstra distance d. Eleven free parameters $a, b, c, d, a', b', c', d', e, f, g$ from three functions m, n, ϕ (left) are used to reconstruct the joint reward function r with 16 fixed values depending on the task (right). The free parameters are defined as $a := m(0,0), b := m(0,1), c := m(1,0), d := m(1,1), a' := m(0,0), b' := m(0,1), c' := m(1,0), d' := m(1,1), e := \phi(d=0), f := \phi(d=1), g := \phi(d=2)$. This constructs an overdetermined linear system with 11 unknowns and 16 constrains. Lower: to make this more intuitive, we could reshape the unknowns as a column vector β and reshape the fixed value functions as column vector y. Then design matrix X is then constructed based on the relationship between agents' locations and the geometry of the grid environment. An optimal solution of β could be constructed as the ordinary least square (OLS) solution of this linear system: $\hat{\beta} = (X^T X)^{-1} X^T y$.

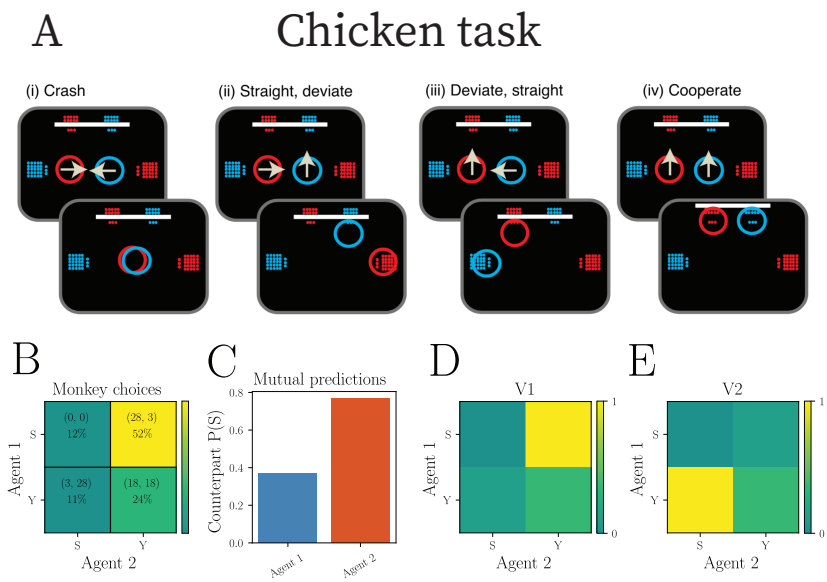


Figure S2: **Monkey's mutual predictions are related to their social hierarchy in a non-cooperative task.** (A) Two monkeys (red and blue circles) choose to go Straight (S) or Yield (Y) in a non-cooperative chicken task. (adapted from [27]) (B) The choices of a pair of monkeys from the experiment (Fig. 2A in [27]). 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.