Theory of Minds: Understanding Behavior in Groups through Inverse Planning

Michael Shum*
Brain and Cognitive Sciences
MIT
mshum@mit.edu

Max Kleiman-Weiner*
Brain and Cognitive Sciences
MIT
maxkw@mit.edu

Michael L. Littman
Computer Science
Brown University
mlittman@cs.brown.edu

Joshua B. Tenenbaum
Brain and Cognitive Sciences
MIT
jbt@mit.edu

Abstract

Human social behavior is structured by relationships. We form teams, groups, tribes, and alliances at all scales of human life. These structures guide multi-agent cooperation and competition, but when we observe others these underlying relationships are typically unobservable and hence must be inferred. Humans make these inferences intuitively and flexibly, often making rapid generalizations about the latent relationships that underlie behavior from just sparse and noisy observations. Rapid and accurate inferences are important for determining who to cooperate with, who to compete with, and how to cooperate in order to compete. Towards the goal of building machine-learning algorithms with human-like social intelligence, we develop a generative model of multi-agent action understanding based on a novel representation for these latent relationships called Composable Team Hierarchies (CTH). This representation is grounded in the formalism of stochastic games and multi-agent reinforcement learning. We use CTH as a target for Bayesian inference yielding a new algorithm for understanding behavior in groups that can both infer hidden relationships as well as predict future actions for multiple agents interacting together. Our algorithm rapidly recovers an underlying causal model of how agents relate in spatial stochastic games from just a few observations. The patterns of inference made by this algorithm closely correspond with human judgments and the algorithm makes the same rapid generalizations that people do.

Introduction

Cooperation enables people to achieve together what no individual would be capable of on her own. From a group of hunters coordinating their movements to an ad-hoc team of programmers working on an open source project, the scale and scope of human cooperation behavior is unique in the natural world (Tomasello 2014; Henrich 2015). However, cooperation exists in a competitive world and finding the right balance between cooperation and competition is a fundamental challenge for anyone in a diverse multi-agent world. At the core of this challenge is figuring out who to cooperate with. How do we distinguish between friend and foe? How can we parse a multi-agent world into groups, tribes, and alliances? Typically when we observe behavior we only get information about these latent relationships sparsely and indirectly through the actions chosen by agents. Furthermore, these inferences are fundamentally challenging because of their inherent ambiguity; we are friend to some and foe to others (Galinsky and Schweitzer 2015). They are also compositional and dynamic; we may cooperate with some agents in order to better compete against another. In order for socially aware AI systems to be capable of acting as our cooperative partners they must learn the latent structure that governs social interaction.

In some domains like sports and formal games, this social structure is known in advance and is essentially written into the environment itself e.g., "the rules of the game" (Kitano et al. 1997; Jaderberg et al. 2018). In contrast we focus on cases where cooperation is more ambiguous or could even be ad-hoc. In real life, people rarely play the same game twice and have to figure out the rules as they go along whether it's the "rules of war" or navigating office politics.

Even young children navigate this uncertainty frequently and display spontaneous cooperation in novel situations from an early age (Warneken and Tomasello 2006; Hamlin, Wynn, and Bloom 2007; Hamann et al. 2011). There is increasing evidence that this early arising ability to do social evaluation and inference relies on "Theory-of-Mind" (ToM) i.e., a generative model of other agents with mental states of their own (Spelke and Kinzler 2007; Kiley Hamlin et al. 2013). People use these models to simulate what another agent might do next or consider what they themselves would do hypothetically in a new situation. From the perspective of building more socially sophisticated machines, ToM acts as a strong inductive bias for predicting actions. Rather than learning statistical patterns of low-level behavior which are often particular to a specific context (e.g., Bob often goes left, then up), an approach based on human ToM constrains inference to behaviors that are consistent with a higher-level mental state such as a goal, a belief or even a false-belief (e.g., Bob likes ice cream). When inference is carried out over these higher-level mental states, the inferences made are more likely to generalize to new contexts in a human-like way (Baker, Saxe, and Tenenbaum 2009; Baker et al. 2017).

Inspired by this ability, we aim to develop a new algorithm that applies these human-like inductive biases towards

*equal contribution

understanding groups of agents in mixed incentive (cooperative / competitive) contexts. These algorithms also serve as models of human cognition that can give us a deeper understanding of human social intelligence. Practically, by capturing the intuitive inferences that people make, our algorithm is more likely to integrate with humans since it will behave in predictable and understandable ways. Our approach builds on two major threads in the literature: generative models for action understanding and game theoretic models of recursive reasoning. Our contribution is the development of a new representation, Composable Team Hiearchies (CTH) for reasoning about how one agent's planning process depends on another and can flexibly capture the kinds of teams and alliances that structure group behavior. We propose that an algorithm using CTH as an inductive bias for Bayesian inverse planning will have the flexibility to represent many types of group plans but is also constrained enough that it will enable the kinds of rapid generalizations that people do. We validate this hypothesis with two behavioral experiments where people are given the same scenarios as the algorithm and are asked to make the same inferences and predictions that the algorithm did.

Related Work

Inferring the latent mental states of agents (e.g., beliefs, desires, and intentions) from behavior features prominently in machine learning and cognitive science (see Albrecht and Stone (2017) for a recent and comprehensive review from the machine learning point of view and Jara-Ettinger et al. (2016) for a developmental perspective). Previous computational treatments similar in spirit to the approach here have focused on making inferences about other individuals acting in a single agent setting (Ng and Russell 2000; Baker, Saxe, and Tenenbaum 2009; Ramırez and Geffner 2011; Evans, Stuhlmüller, and Goodman 2016; Nakahashi, Baker, and Tenenbaum 2016; Baker et al. 2017; Rabinowitz et al. 2018). When these tools are applied to multi-agent and game theoretic contexts they have focused on dyadic interactions (Yoshida, Dolan, and Friston 2008; Ullman et al. 2009; Kleiman-Weiner et al. 2016; Raileanu et al. 2018). Dyadic interactions are significantly simpler from a representational perspective since an observer must merely determine whether each agent is cooperating or competing.

However, when the number of agents increases beyond a two player dyadic interaction, the problem of balancing cooperation and competition often takes on a qualitatively different character. Going from two to three or more players means the choice is no longer whether to simply cooperate or compete. Instead agents must reason about which agents they should cooperate with and which they should compete with. In a dyadic interaction there is no possibility of cooperating *to* compete or the creation of more complicated alliances and groups.

Computational Formalism

Stochastic Games

We study multi-agent interactions in stochastic games which generalize single-agent Markov Decision Processes to sequential decision making environments with multiple agents. Formally, a stochastic game, G, is the tuple $\langle n, \mathcal{S}, \mathcal{A}_{1...n}, T, R_{1...n}, \gamma \rangle$ where n is the number of agents, $\mathcal{S}$ is a set of states, $\mathcal{A}_{1...n}$ is the joint action space with $\mathcal{A}_i$ the set of actions available to agent i, $T(s, a_{1...n}, s')$ is the transition function which describes the probability of transitioning from state s to s' after $a_{1...n}$, $R_{1...n}(s, a_{1...n}, s')$ is the reward function for each agent, and γ is the discount factor (Bowling and Veloso 2000; Filar and Vrieze 2012). The behavior of each agent is defined by a policy $\pi_{1...n}(s)$ which is the probability distribution over actions that each agent will take in state s.

There are many different notions of what it means to "solve" a stochastic game. Many of these concepts rely on notions of finding a best-response (Nash) equilibrium (Littman 1994; 2001; Hu and Wellman 2003). While solution concepts based on equilibrium analyses provide some constraints on the policies agents will use, they cannot provide a way to explain behavior carried out by bounded or cooperative agents who are willing to play a dominated strategy to help another agent. When games are repeated, there are often an explosion of equilibrium and these methods do not provide a clear method for choosing between them. Finally, there is ample evidence that both human behavior and judgments are not well explained by equilibrium thinking (Wright and Leyton-Brown 2010). On the other hand, without constraints from rational planning on the types of policies that agents are expected to use, there will be no way for an observer to generalize or predict how an agent's policy will adapt to a new situation or context that the agent has not been observed to act in.

Group Plan Representation

In this section we build up a representation for multi-agent interaction that can be used to compute policies for agents in novel situations, but is also sufficiently constrained that it can be used for rapid inference. We first introduce two simple planning operators based on individual best-response (`BR`) and joint-planning (`JP`). We then show how they can be composed together using a `REPLACE` operator into Composable Team Hierarchies (CTH) which enable the flexible representation of teams and alliances within a MARL context.

Operator Composition (`REPLACE`) We first define the `REPLACE` operator that takes an N player stochastic game, G and a policy π_R indexed to a particular set of players (R) and returns an $N - |R|$ agent stochastic game G' with the agents in R removed. This new game G' embeds π_R in G to generate dynamics for R such that from the perspective of an agent in G', the agents in R are now stationary parts of the environment in G' predictable from their policies. Formally $\texttt{REPLACE}(G, \pi_R)$ creates a game G' identical to G but with a reduced action space that excludes the agents in R and modifies the transition function as follows:

$$T_{G'}(s'|s, a_{-R}) = \sum_{a_R} T_G(s'|s, a_{-R}, a_R) \prod_{r \in R} \mathbb{1}(\pi_r(s) = a_r)$$

where the $-R$ refers to all agents other than those in R.

Best Response (`BR`) The best response operator `BR` takes a game G with a single agent i and returns a policy π_i for

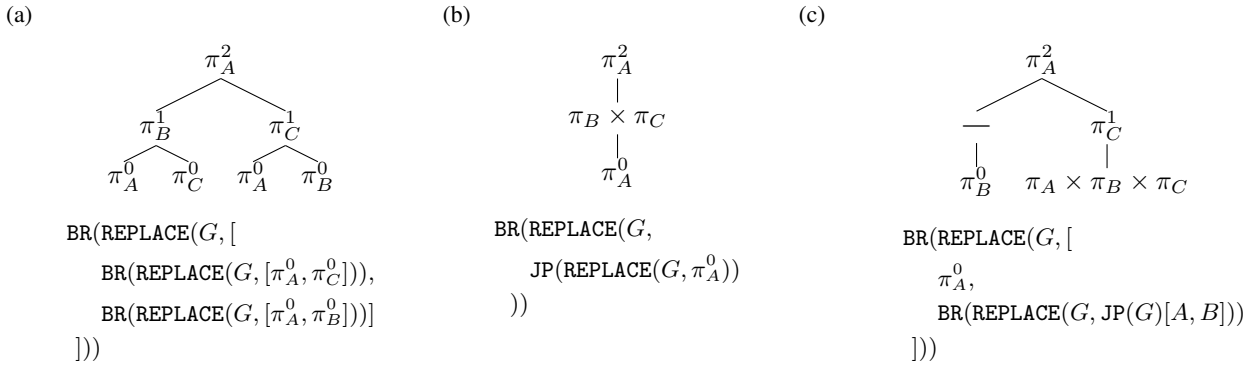

Figure 1: Example composition of base policies and operators to construct different types of teams with agents of variable sophistication. See text for descriptions of the operators and descriptions of these models.

just that agent. Planning here is defined through the standard Bellman equations.

$$Q(s, a_i) = \sum_{s'} T(s'|s, a_i)[R_i(s, a_i, s') + \gamma \max_{a_i'} Q(s', a_i')]$$

$$\pi_i(s) = \arg\max_{a_i} Q(s, a_i)$$

where ties are broken with uniform probability.

Joint Planning (JP) A second planning operator JP generates cooperative behavior through joint planning. Each individual agent considers itself part of a hypothetical centralized "team-agent" that has joint control of all the agents that are included in the team and optimizes the joint reward of that team (Sugden 2003; Bratman 2014). Planning under this approach combines all the agents in a game into a single agent and finds a joint plan which optimizes that group objective (De Cote and Littman 2008; Oliehoek, Spaan, and Vlassis 2008; Kleiman-Weiner et al. 2016). If J is the set of agents that have joined together as a team, their joint-plan can be characterized as:

$$Q^J(s, a_J) = \sum_{s'} T(s'|s, a_J) * \Big(\sum_{j \in J} R_j(s, a_j, s') + \gamma \max_{a_J'} Q^J(s', a_J') \Big)$$

Each agent plays its role in the team plan by marginalizing out the actions of all the other agents: $\pi_i(s) = \arg\max_a Q^J(s, a)$ where ties are broken with uniform probability. Thus JP takes a $N > 1$ agent game G as input and returns policies π_J as if all agents in G are cooperating with each other towards a joint goal.

Base Policies π^0 Base policies (π^0) are non-strategic i.e., they take actions independent of the other agents. A common choice is to act randomly or to choose a locally optimal option ignoring all other agents. See (Wright and Leyton-Brown 2014) for the various choices one could make about the level-0 policy in matrix-form games, some of which could be extended to stochastic games.

Composable Team Hierarchies

We now show how with just these three simple operators (REPLACE, JP, BR) and a set of base policies ($\pi^0_{1\ldots n}$) we can create complex team plans that vary in both their team structures as well as their sophistication. We start by noting that the output of JP and BR (policies) is an input to REPLACE, and the output of REPLACE (games with fewer players) is an input to JP and BR. When composed together, these operators generate hierarchies of policies.

Figure 1 shows how these planning procedures can be composed together to create strategic agents (using BR), teams of cooperative agents (using JP) and compositional combinations of the two. Even with just three players there are a combinatorial number of possible team partitions (all playing together, two against one, no teams) and higher and lower levels of sophisticated agents within those partitions. When shown hierarchically, this representation mirrors the tree-like structure of a grammar producing an "infinite use of finite means" – the key benefit of a composable representation. We call these structures Composable Team Hierarchies (CTH). We now show how previous approaches from the literature can be subsumed under CTH, which unifies some previous approaches and enables new ways of reasoning about plans.

Level-K Planning One technique common in behavioral game theory is iterative best response which is often called level-K or cognitive hierarchy (Camerer, Ho, and Chong 2004; Wright and Leyton-Brown 2010). In brief, an agent operating at level K assumes that other agents are using $K-1$ level policies. This approach has also been extended to sequential decision making in reinforcement learning (Yoshida, Dolan, and Friston 2008; Kleiman-Weiner et al. 2016; Lanctot et al. 2017). By considering only a finite number (K) of these best responses, an infinite regress is prevented. This also captures some intuitive constraints on bounded thinking. This approach to multi-agent planning is to replace all other agents with a slightly less strategic $k-1$ policy. Importantly, this formalism maps a multi-agent planning problem into a hierarchy of nested single-agent planning problems. This recursive hierarchy grounds out in

level-0 models (π_i^0) which we described above as base policies.

Figure 1a shows how a level-K policy with $K = 2$ for agent A can be constructed by iterating between the BR and REPLACE operators. The CTH shows a level-2 A best responding to level-1 models of B and C who are best responding to level-0 base policies of A & C and A & B respectively.

Cooperative Planning While Level-K representations can capture certain aspects of strategic thinking i.e., how to best respond in one's own interest to other agents, it is not sufficient to generate the full range of social behavior. Specifically it will not generate cooperative behavior when cooperation is dominated in a particular scenario. However cooperative behavior between agents that form teams and alliances is commonly observed. An agent may be optimizing for a longer horizon where the game is repeated or one's reputation is at stake. Furthermore, certain agents may have intrinsic pro-social dispositions and an observer must be able to reason about these. A cooperative stance towards a problem can be modeled as a DEC-MDP (Oliehoek, Spaan, and Vlassis 2008). In CTH this stance is easily represented. For instance starting with a base game G that has players (A, B, C) one can compute all three policies for working together as: $\texttt{JP}(G)$.

Composing Cooperation and Competition In addition to these two well studied formalisms, CTH can represent a range of possible social relationships that are not expressible with level-K planning or cooperative planning alone. Figure 1b combines both operators to describe a cooperate *to* compete stance. Under this CTH A best responds to B & C cooperating to compete against a naive version of A's own behavior. Figure 1c depicts agent A best responding to both a naive B and model of C that is acting to betray the group of three. The CTH representation can capture an A which is acting to counter a perceived betrayal by C. These examples show the representational flexibility of CTH and its ability to intuitively capture different social stances that agents might have towards each other.

Inverse Group Planning

Observers can use CTH to probabilistically infer the various stances that each agent takes towards the others. Agents represent their uncertainty over the CTH for agent i as $P(\texttt{CTH}_i)$, their prior beliefs before seeing any behavior. These beliefs can be updated in response to the observation of new behavioral data using Bayes rule:

$$P(\texttt{CTH}_i|\mathbf{s}, \mathbf{a_i}) \propto P(\texttt{CTH})P(\mathbf{a_i}|\mathbf{s}, \texttt{CTH}_i) \quad (1)$$

$$= P(\texttt{CTH}_i)\prod_{t=1}^{T} P(a_{i,t}|s_t, \texttt{CTH}_i) \quad (2)$$

where $\mathbf{s}$ and $\mathbf{a_i}$ are sequences of states and actions from time $1 \dots T$. $P(a_{i,t}|s_t|\texttt{CTH}_i)$ is the probability of a given action under the induced hierarchy of goal-directed planning as determined by a given CTH. We use the Luce-choice decision rule to transform the Q-values of each action under planning into a probability distribution:

$$P(a_{i,t}|s_t, \texttt{CTH}_i) \propto \exp(\beta * Q^*_{\texttt{CTH}}(s, a)) \quad (3)$$

where β controls the degree to which the observer believes agents are able to correctly maximize their future expected utility at each time step. When $\beta \to \infty$ the observer believes that agents are perfect optimizers, as $\beta \to 0$ the observer believes the other agents are acting randomly. $Q^*_{\texttt{CTH}}(s, a)$ are the optimal Q-values of the root agent in a given CTH.

In theory, the number of CTH considered by an observer could be infinite since the number of levels in the hierarchy does not have to be bounded. As this would make inference impossible, a realistic assumption is to assume some maximum level of sophistication which bounds the number of levels in the hierarchy (Yoshida, Dolan, and Friston 2008; Kleiman-Weiner et al. 2016). Another possibility is to put a monotonically decreasing probability distribution on larger CTH as is done in the cognitive hierarchy model. Finally, since we have posed this IRL problem as probabilistic inference, Markov Chain Monte Carlo (MCMC) algorithms and other sampling approaches might enable the computation of $P(\texttt{CTH}_i|\mathbf{s}, \mathbf{a_i})$ even when the number of hypothetical CTH are large. In this work we are agnostic to the how the polices are computed as any reinforcement learning algorithm is possible. In our simulations we used a simple version of Monte Carlo Tree Search (MCTS) based on Upper Confidence Bound applied to Trees (UCT) which selectively explores promising action sequences (Browne et al. 2012).

Experiments

Our two experiments were carried out in a spatial stag-hunt domain. In these scenarios, the agents control hunters who gain rewards by capturing prey: stags and hares. The hunters and stags can move in the four cardinal directions and can stay in place. The hares do not move but the stags move away from any nearby hunters so must be cornered to be captured. All moves happen simultaneously and agents are allowed to occupy the same squares.

Hunters are rewarded for catching either type of prey by moving into its square, and when any prey is caught the game terminates. Catching a hare is worth one point to the hunter who catches it. Catching a stag requires coordination but hunters split 20 points when capturing it. At least two hunters must simultaneously enter the square with the stag on it to earn the points. The stag-hunt is a common example used to demonstrate the challenges of coordinating on joint action and the creation of cooperative cultural norms (Skyrms 2004). Previous work on the spatial stag-hunt has mostly focused on modeling behavior, not inferences, in two-player versions. As noted before, with only two hunters there are limited possibilities for different team arrangements (Yoshida, Dolan, and Friston 2008; Peysakhovich and Lerer 2018).

We designed nine different scenes in this stag-hunt world that each had three hunters, two stags, and two hares (Figure 2). Different scenes had different starting positions and different spatial layouts which means that both people and our algorithm must generalize across different environments

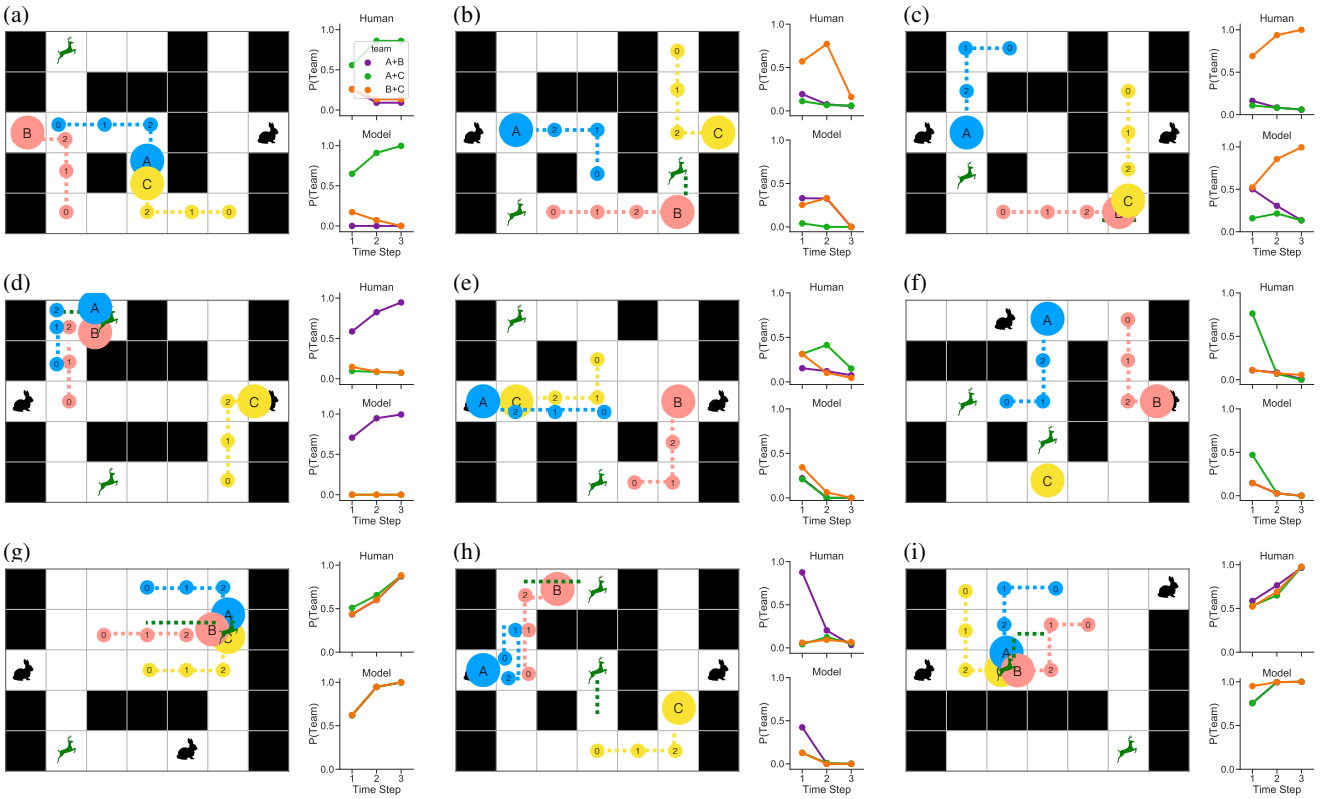

Figure 2: Experimental scenarios and results for Experiment 1. Each scene involved 3 time steps from a starting state. The hunters were represented with circles. Each agent's movement path is traced out with dotted lines with small numbered dots indicating the position in a given time step. Human participants saw a more dynamic scene that played out over the course of the experiment instead of seeing the full trajectory at once. To the right of each scenario are plots showing the average human inferences (top) and Bayesian model average results run on the same scene that participants saw.

and contexts. This makes it less likely that a heuristic approach based on features will work. Instead we can test whether people invert an underlying group planning process. In both experiments (N=37, all USA), each participant was tested on all nine scenarios. Participants watched the scene unfold a few moves at a time. All human data was collected on Amazon Mechanical Turk. In Experiment 1 we compare our algorithm against the inferences people made about the underlying structure i.e., who is cooperating with who. In Experiment 2 we compare our algorithm against people's ability to predict the next action in a scene. In both experiments we compare human participant data with our Bayesian inference with a uniform prior over depth-1 CTH. When modeling human judgments and behavior β usually corresponds to the noise in the utility maximization process where non-optimal decisions are made proportional to how close their utility is to optimal. In the team inference experiment (experiment 1) participants used a continuous slider to report their judgments so $\beta = 1$ was used for model comparison while in the action-prediction experiment (experiment 2) subjects made a discrete choice to report their predictions so a higher $\beta = 5$ was used.

Experiment 1: Team Inference

For each scenario, participants made judgments about whether A&B, B&C, and C&A were cooperating at three different time points by selecting a range between 0 and 100 on a slider. These ratings were averaged together and normalized to 0 and 1. In total, we collected a total of 81 distinct data points. Figure 2 shows the human data and model results for each of the nine scenarios at each time point. We did not find any systematic individual differences among human participants.

The model performs well across all inferred team structures: when all three players are on the same team (g, i), when just two are working together (a, c, d) and when all three are working independently (b, e, f, h). These inferences were made based on information about the abstract team structure since they were made before any of the actual outcomes were realized. Even a single move was often sufficient to give away the collective intentions of the group of hunters. Finally, the model also handles interesting reversals. In situation (b), one might infer that B and C were going to corner the stag but this inference is quickly overturned when C goes for a hare instead at the last minute. Situation (h) also contains a change of mind. At first it seems A is following B

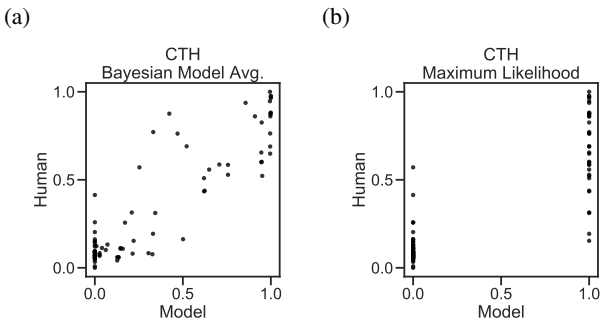

Figure 3: Quantification of algorithm inferences in Experiment 1 compared to human judgments. (a) Bayesian model averaging explains a high degree of variance in human judgments. (b) The maximum likelihood CTH captures some of the coarse grained aspects of the inferences but does not capture the uncertainty in people's judgments.

to capture the stag together but he then reverses and goes for the hare. Just this single reversal was sufficient to flip people's judgments about the underlying team structure and our algorithm captures this.

There were also a few circumstances where people's judgments significantly deviated from our algorithm. For instance, in scenario (c) people were quicker to infer that B and C are on the same team while the model has greater uncertainty. This difference might reflect the fact that people put a higher prior on cooperative CTH while we used a uniform prior. Indeed, increasing the prior on cooperative CTH results in more human-like inferences for this scenario. Most of the differences were more subtle. The model makes stronger predictions (closer to 0 and 1) while people integrated information more slowly and less confidently. Figure 3 and Table 1a show a quantification of how well our algorithm can act as a model for human judgments. Bayesian model averaging across the uncertainty in the underlying CTH (Figure 3a) did a better job of capturing human team inferences than did the CTH with the highest likelihood (Figure 3b). Thus a full Bayesian approach seems to be needed to capture the nuanced and graded judgments that people make as they integrate over the ambiguous behavior of the agents.

Experiment 2: Action Prediction

In a second experiment using the same set of stimuli as Experiment 1 we compared our system's ability to predict the next action with people's predictions. Each participant was given the choice to select the action (from those available) for each of the 3 hunters. Averaging over the participants gives a distribution over the next action for human participants. Across all nine scenarios, we elicited 53 judgements which generated 216 distinct data points. Since our computational formalism is a generative model the same algorithms that was used for team inferences is also tested on action predictions.

Figure 4 and Table 1b show the ability of this algorithm to predict the human action prediction distribution. We find

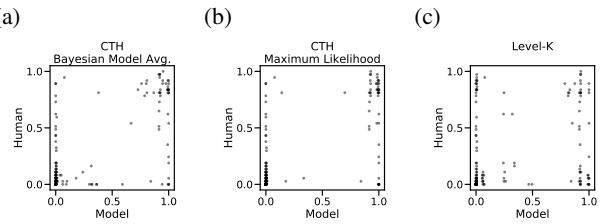

Figure 4: Quantification of algorithm predictions in Experiment 2 compared to human predictions. Both (a) Bayesian model averaging and (b) the maximum likelihood CTH explains a high degree of variance in human predictions. (c) Level-K models explain significantly less variance in human predictions.

(a) Experiment 1

	BMA-CTH	MLE-CTH
R	**0.90**	0.86
RMSE	**0.18**	0.28

(b) Experiment 2

	BMA-CTH	ML-CTH	Level-K
R	**0.74**	0.73	0.38
RMSE	**0.27**	0.28	0.41

Table 1: Pearson correlation coefficients (R) and root mean square error (RMSE) for both experiments. Higher R and lower RMSE indicate explaining more of the variation in the human (a) judgments and (b) predictions. BMA is Bayesian model average and ML is maximum likelihood.

a relatively high-correlation ($R > 0.7$) but we do not find as large of a difference between the Bayesian model averaging over CTH and the maximum likelihood CTH. This is likely due to the fact that each participant directly selected the action they thought was most likely rather than give a graded measure of confidence. Still both of these CTH based models out-perform a Level-K model which does not allow for the use of the JP operator.

Discussion

Our contribution is a novel representation for extending single-agent generative models of action understanding to the richness of multi-agent interaction. In human cognition, the ability to infer the mental states of others is often called Theory-of-Mind and here we develop a Theory-of-Minds which explains and predicts group behavior. The core of this work is to build on two key insights from how people learn in general and in particular learn about other people (Tenenbaum et al. 2011):

1. Agents have the ability to construct generative models of other agents and use those models to reason about future actions by simulating their planning. With models of other agents and a model of the environment, agents can predict what will happen next through forward simulation. With these future-oriented predictions of what other agents will do, individuals can better generate their own plans. They

can even use these models hypothetically in order to predict what an agent would do, or counterfactually to predict what another agent would have done.

2. The inferences people make about others take place at a high level of abstraction. For instance, people learn about who is cooperating with whom and why, rather than reasoning directly about the likelihood of a particular sequence of actions in a specific situation. While in some sense, these abstract inference are more complex, they drastically reduce a hypotheses space about every action an agent might take to a much smaller hypothesis space of actions that serve a social purpose. These abstract reasons generalize in ways that mere patterns of behavior do not.

In this work, we took inspiration from the ways that human observers think abstractly about alliances, friendships, and groups. We formalized these concepts in a multi-agent reinforcement learning formalism and used them as priors to make groups tractably understandable to an observer. Our model explains the fine-grained structure of human judgment and closely matches the predictions made by human observers in a novel and varying three-agent task.

There are still many avenues for future work. While the approach described here can work well for small groups of agents, the computations involved scale poorly with the number of agents. Indeed, when interacting with a large number of agents more coarse-grained methods which ignore individual mental states might be required (Yang et al. 2018). Another way forward is to constrain the possible types of CTH to consider. For instance, when dealing with a large number of agents, people seem to use group membership cues, some of which are directly observable such as style of dress or easily inferable such as language spoken (Liberman, Woodward, and Kinzler 2017). These cues could rapidly prune the number of CTHs considered but also could lead to biases. Another possible route to scaling these methods is through sophisticated state abstractions such as those in deep multi-agent reinforcement learning where agents are trained for cooperation and competition (Leibo et al. 2017; Perolat et al. 2017; Lowe et al. 2017; Lerer and Peysakhovich 2017; Peysakhovich and Lerer 2018; Foerster et al. 2018). Self-play and feature learning based methods might also be useful for generating interesting base policies to build on in our CTH representation (Hartford, Wright, and Leyton-Brown 2016).

Our current set of experiments looked at situations where team coordination required spatial convergence. Future work will look at environments with buttons that can open and close doors or by giving agents the ability to physically block others. In these scenarios heuristics based on spatial convergence will not correlate with human judgments and higher order CTH may be needed to identify the underlying team structures. Finally, endowing multi-agent reinforcement learning agents with the ability to do CTH inference could give these agents the ability to more effectively reason about and coordinate with others.

While we propose a method of understanding and planning with agents that have known teams, agents are frequently in scenarios where team structures have yet to be established e.g., the first day of kindergarten. In future work, we hope to explore how agents can identify the best teammates in an environment and create a social relationship with them. Based on previous actions an observer could begin to predict an agent's likelihood of changing its social stance through changes in the CTH structure. Finally, social norms and anti-social behavior such as punishing and disliking are not easily captured in the current version of the CTH representation. Future work will extend CTH with new operators that expand its flexibility.

Acknowledgments This work was supported by a Hertz Foundation Fellowship, the Center for Brains, Minds and Machines (CBMM), NSF 1637614, and DARPA Ground Truth. We thank Mark Ho for comments on the manuscript and for helping to develop these ideas.

References

Albrecht, S. V., and Stone, P. 2017. Autonomous agents modelling other agents: A comprehensive survey and open problems. *arXiv preprint arXiv:1709.08071*.

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

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

Bowling, M., and Veloso, M. 2000. An analysis of stochastic game theory for multiagent reinforcement learning. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Bratman, M. E. 2014. *Shared agency: A planning theory of acting together*. Oxford University Press.

Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; and Colton, S. 2012. A survey of monte carlo tree search methods. *Computational Intelligence and AI in Games, IEEE Transactions on* 4(1):1–43.

Camerer, C. F.; Ho, T.-H.; and Chong, J.-K. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics* 861–898.

De Cote, E. M., and Littman, M. L. 2008. A polynomial-time nash equilibrium algorithm for repeated stochastic games. In *24th Conference on Uncertainty in Artificial Intelligence*. Citeseer.

Evans, O.; Stuhlmüller, A.; and Goodman, N. D. 2016. Learning the preferences of ignorant, inconsistent agents. In *AAAI*, 323–329.

Filar, J., and Vrieze, K. 2012. *Competitive Markov decision processes*. Springer Science & Business Media.

Foerster, J.; Chen, R. Y.; Al-Shedivat, M.; Whiteson, S.; Abbeel, P.; and Mordatch, I. 2018. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 122–130. International Foundation for Autonomous Agents and Multiagent Systems.

Galinsky, A., and Schweitzer, M. 2015. *Friend and Foe: When to Cooperate, when to Compete, and how to Succeed at Both*. Random House.

Hamann, K.; Warneken, F.; Greenberg, J. R.; and Tomasello, M. 2011. Collaboration encourages equal sharing in children but not in chimpanzees. *Nature* 476(7360):328–331.

Hamlin, J. K.; Wynn, K.; and Bloom, P. 2007. Social evaluation by preverbal infants. *Nature* 450(7169):557–559.

Hartford, J. S.; Wright, J. R.; and Leyton-Brown, K. 2016. Deep learning for predicting human strategic behavior. In *Advances in Neural Information Processing Systems*, 2424–2432.

Henrich, J. 2015. *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hu, J., and Wellman, M. P. 2003. Nash q-learning for general-sum stochastic games. *Journal of machine learning research* 4(Nov):1039–1069.

Jaderberg, M.; Czarnecki, W. M.; Dunning, I.; Marris, L.; Lever, G.; Castaneda, A. G.; Beattie, C.; Rabinowitz, N. C.; Morcos, A. S.; Ruderman, A.; et al. 2018. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *arXiv preprint arXiv:1807.01281*.

Jara-Ettinger, J.; Gweon, H.; Schulz, L. E.; and Tenenbaum, J. B. 2016. The naïve utility calculus: computational principles underlying commonsense psychology. *Trends in cognitive sciences* 20(8):589–604.

Kiley Hamlin, J.; Ullman, T.; Tenenbaum, J.; Goodman, N.; and Baker, C. 2013. The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental science* 16(2):209–226.

Kitano, H.; Tambe, M.; Stone, P.; Veloso, M.; Coradeschi, S.; Osawa, E.; Matsubara, H.; Noda, I.; and Asada, M. 1997. The robocup synthetic agent challenge 97. In *Robot Soccer World Cup*, 62–73. Springer.

Kleiman-Weiner, M.; Ho, M. K.; Austerweil, J. L.; Littman, M. L.; and Tenenbaum, J. B. 2016. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Lanctot, M.; Zambaldi, V.; Gruslys, A.; Lazaridou, A.; Perolat, J.; Silver, D.; Graepel, T.; et al. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, 4193–4206.

Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 464–473. International Foundation for Autonomous Agents and Multiagent Systems.

Lerer, A., and Peysakhovich, A. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*.

Liberman, Z.; Woodward, A. L.; and Kinzler, K. D. 2017. The origins of social categorization. *Trends in cognitive sciences* 21(7):556–568.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *ICML*, volume 94, 157–163.

Littman, M. L. 2001. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, 322–328.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, O. P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 6382–6393.

Nakahashi, R.; Baker, C. L.; and Tenenbaum, J. B. 2016. Modeling human understanding of complex intentional action with a bayesian nonparametric subgoal model. In *AAAI*, 3754–3760.

Ng, A. Y., and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 663–670. Morgan Kaufmann Publishers Inc.

Oliehoek, F. A.; Spaan, M. T.; and Vlassis, N. 2008. Optimal and approximate q-value functions for decentralized pomdps. *Journal of Artificial Intelligence Research* 32:289–353.

Perolat, J.; Leibo, J. Z.; Zambaldi, V.; Beattie, C.; Tuyls, K.; and Graepel, T. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, 3646–3655.

Peysakhovich, A., and Lerer, A. 2018. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2043–2044. International Foundation for Autonomous Agents and Multiagent Systems.

Rabinowitz, N. C.; Perbet, F.; Song, H. F.; Zhang, C.; Eslami, S.; and Botvinick, M. 2018. Machine theory of mind. In *Proceedings of the 35nd International Conference on Machine learning*.

Raileanu, R.; Denton, E.; Szlam, A.; and Fergus, R. 2018. Modeling others using oneself in multi-agent reinforcement learning. In *Proceedings of the 35nd International Conference on Machine learning*.

Ramırez, M., and Geffner, H. 2011. Goal recognition over pomdps: Inferring the intention of a pomdp agent. In *IJCAI*, 2009–2014. IJCAI/AAAI.

Skyrms, B. 2004. *The stag hunt and the evolution of social structure*. Cambridge University Press.

Spelke, E. S., and Kinzler, K. D. 2007. Core knowledge. *Developmental science* 10(1):89–96.

Sugden, R. 2003. The logic of team reasoning. *Philosophical explorations* 6(3):165–181.

Tenenbaum, J. B.; Kemp, C.; Griffiths, T. L.; and Goodman, N. D. 2011. How to grow a mind: statistics, structure, and abstraction. *Science* 331(6022):1279.

Tomasello, M. 2014. *A natural history of human thinking*. Harvard University Press.

Ullman, T.; Baker, C.; Macindoe, O.; Evans, O.; Goodman, N.; and Tenenbaum, J. B. 2009. Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems*, 1874–1882.

Warneken, F., and Tomasello, M. 2006. Altruistic helping in human infants and young chimpanzees. *Science* 311(5765):1301–1303.

Wright, J. R., and Leyton-Brown, K. 2010. Beyond equilibrium: Predicting human behavior in normal-form games. In *AAAI*.

Wright, J. R., and Leyton-Brown, K. 2014. Level-0 meta-models for predicting human behavior in games. In *Proceedings of the fifteenth ACM conference on Economics and computation*, 857–874. ACM.

Yang, J.; Ye, X.; Trivedi, R.; Xu, H.; and Zha, H. 2018. Deep mean field games for learning optimal behavior policy of large populations. In *International Conference on Learning Representations*.

Yoshida, W.; Dolan, R. J.; and Friston, K. J. 2008. Game theory of mind. *PLoS Computational Biology* 4(12).