Policy Communication for Coordination with Unknown Teammates

Trevor Sarratt and **Arnav Jhala**
University of California Santa Cruz
{tsarratt, jhala}@soe.ucsc.edu

Abstract

Within multiagent teams research, existing approaches commonly assume agents have perfect knowledge regarding the decision process guiding their teammates' actions. More recently, ad hoc teamwork was introduced to address situations where an agent must coordinate with a variety of potential teammates, including teammates with unknown behavior. This paper examines the communication of intentions for enhanced coordination between such agents. The proposed decision-theoretic approach examines the uncertainty within a model of an unfamiliar teammate, identifying policy information valuable to the collaborative effort. We characterize this capability through theoretical analysis of the computational requirements as well as empirical evaluation of a communicative agent coordinating with an unknown teammate in a variation of the multiagent pursuit domain.

Introduction

Coordinating a team of autonomous agents is a challenging problem. Agents must act in such a way that progresses toward the achievement of a goal while avoiding conflict with their teammates. In information asymmetric domains, it is often necessary to share crucial observations in order to collaborate effectively. In traditional multiagent systems literature, these teams of agents share an identical design for reasoning, planning, and executing actions, allowing perfect modeling of teammates. Ad hoc teamwork (Stone et al. 2010) further complicates this problem by introducing a variety of teammates with which an agent must coordinate. In these scenarios, one or more agents within a team can be unfamiliar, having unknown planning capabilities guiding their behavior.

Much of the existing ad hoc teamwork research focuses on reinforcement learning and decision theoretic planning. Agents use models of known behavior to predict an ad hoc agent's actions, using decision theory to maximize expected utility in instances where the predicted actions are uncertain (Barrett, Stone, and Kraus 2011). Online learning refines these models with observations of behaviors during execution, increasing the accuracy of the models' predictions, permitting the team to coordinate more effectively (Barrett et al. 2012; 2013). Though such approaches typically assume teammates retain a static model of behavior for the duration of the task, alternate belief revision techniques have been shown to be effective when coordinating with inconsistent agents (Sarratt and Jhala 2015). A deeper analysis of posterior belief updates and the impact of priors can be found in (Albrecht, Crandall, and Ramamoorthy 2015; Albrecht and Ramamoorthy 2014).

This paper further addresses the problem of planning under teammate behavior uncertainty by introducing the concept of intentional multiagent communication within ad hoc teams. In partially observable multiagent domains, agents much share information regarding aspects of the environment such that uncertainty is reduced across the team, permitting better coordination. Similarly, we consider how communication may be utilized within ad hoc teams to resolve behavioral uncertainty. Transmitting intentional messages allows agents to adjust predictions of a teammate's individual course of action. With more accurate predictions of team behavior, an agent can better select its personal actions to support team cohesion. The main contribution of this paper is a decision-theoretic approach for evaluating intentional communicative acts within a finite planning horizon. In contrast to traditional multiagent communication applications where communicative acts are few in number, we allow the agent to consider the entire set of states encountered when planning. For this consideration, we describe an efficient method of evaluating potential communicative acts. Secondly, we characterize the interaction between learning, communication, and planning. In short, an ad hoc agent coordinating with an unknown teammate can identify uncertainties within its own predictive model of teammate behavior then request the appropriate policy information, allowing the agent to adapt its personal plan. To demonstrate the generality of the approach, this process is evaluated using two types of agents: an agent which learns purely from observation and an agent that updates a belief distribution over a set of known models.

Uncertainty in Team Behavior

For clarity, we adopt the language of Markov decision processes (MDPs), wherein an agent performs actions to tran-

sition between states and is rewarded with various utilities assigned to those states. Formally, we describe the problem as a tuple $\langle S, A, P, R \rangle$, where

- S is a finite set of states.
- A is a finite set of actions.
- T is a transition probability function. $T(s, a, s') = Pr(s'|s, a)$ represents the probability of transitioning to state s' from s when taking action a.
- R is a reward function, where $R(s)$ is the reward obtained by taking an action in some state s' and transitioning to state s.

We note that in many domains, it is more appropriate to consider partially observable models, wherein aspects of the state are not known explicitly, but must be reasoned over using individual observations. For the sake of analyzing only behavioral uncertainty, we omit such considerations, though agents may additionally adopt beliefs for such factors in the appropriate applications.

Consider the space of policies that could describe the behavior of a teammate, where $\pi : S \mapsto A$ is a mapping of actions to states in our domain. Clearly, it is intractable to reason exhaustively over the entire space of policies. In order to reduce the space of possible behaviors, it is common to make assumptions regarding the classes of teammates encountered within a domain. Best response (Agmon and Stone 2012; Stone et al. 2013; Stone, Kaminka, and Rosenschein 2010), greedy (Barrett and Stone 2014), or prior-learned domain-specific models (Barrett, Stone, and Kraus 2011) are used. Regardless of which approach is used, a coordinating agent will have a predictive model of a teammate's actions, M, where $M(s, a)$ is the probability of action a given state s. This model can be composed of predictions from individual types of teammates weighted by an agent's belief distribution over the set(Barrett, Stone, and Kraus 2011; Sarratt and Jhala 2015), a learned probabilistic model provided by reinforcement learning techniques (Albrecht and Ramamoorthy 2012; Barrett et al. 2013), or a mixture of both (Barrett et al. 2013).

Given an agent's predictive model of its team, we can compute the finite horizon expected value of its current policy under the uncertainty of its collaborators' future actions.

$$V_\pi^0(s) = R(s),$$

$$V_\pi^h(s) = R(s) + \sum_{a \in A} Pr(a|s) \left(\sum_{s'} T(s, a, s') V_\pi^{h-1}(s') \right)$$

Here, we use M to supply probabilities of teammate actions while our rational agent maximizes its expected payoff by assigning $Pr(a_i|s) = 1$ where $a_i = max_{a_j \in A} \left(\sum_{s'} T(s, a, s') V(s') \right)$ and $Pr(a_k|s) = 0 \; \forall a_k \neq a_i$. This decision-theoretic approach to planning under uncertainty of teammate behavior allows the agent to act optimally with respect to the information available to it. This is not to say the team will converge to an optimal joint-policy, particularly if the predictive model's probabilities to do not align with the teammates' true policies. Sharing policy information can refine the predictive accuracy of the agent's team model, allowing for policy adjustments monotonically increasing the agent's expected reward.

Communication

Across many communicative multiagent frameworks, such as the COM-MTDP model (Pynadath and Tambe 2002) and STEAM (Tambe 1997), communicative actions are often limited to sharing observations or direct state information (Roth, Simmons, and Veloso 2007). As agents in such systems have complete information regarding the planning capacities of their teammates, they can simply use the shared information to compute precisely how all agents will act. Since the policies of teammates is the source of uncertainty in ad hoc teams, it follows that policy information is a promising target for communicative acts.

In early decision theoretic agent communication literature, various types of communication were theoretically validated in their effect on coordinating multiple agents. These included intentional messages (Gmytrasiewicz, Durfee, and Wehe 1991), questions, proposals, threats, imperatives, and statements of propositional attitudes (Gmytrasiewicz, Durfee, and Rosenschein 1995). In each case, providing or requesting information adjusted one or more agents' policies through refining an agent's expectations of either its own policy's likelihood of success or the intentions of another agent acting within the same environment. Analogously, the refinement of predicted action probabilities and, consequently, an improved policy for a coordinating agent is desirable for ad hoc teams.

Whereas the broadcast of the intention of pursuing a goal addresses multiple state-action pairs within an agent's policy computation, we must consider that an unfamiliar teammate may not possess the capability of reasoning with high level abstractions such as joint plans or hierarchical goal decompositions. However, we put forth the observation that all agents involved are universally tasked with assigning actions to states, independent of the particular planning implementation details. These states are comprised of features which may be embodied in the world or calculated by an agent from the joint history of interaction (Chakraborty and Stone 2013), such as in the case of belief states used in POMDP planners (Pynadath and Tambe 2002). As such, we will consider a single state-action pair as the most granular form of information potentially gained from an intentional communicative act. We leave collections of state-action pairs, imaginably employed for communicating sequences of states/actions or for sharing hierarchical abstractions of policy information, for future work.

From singular state-action pairs, two types of communicative acts are immediately obvious: instructive commands, asserting that a teammate perform a specific action when a state is encountered, and intentional queries, requesting what action the teammate intends to perform at the given state. The former relies on an agent planning for multiple agents within the team as a centralized planner then providing instruction to teammates. As we are primarily interested in learning the behaviors of teammates and coordinating in a decentralized fashion, we will omit this type of communica-

tion and focus instead on gaining information about ad hoc teammates.

The Value of Information

Clearly, having complete knowledge of a teammate's intended behavior would allow optimal planning on the part of the coordinating agent. However, the exchange of such intentions given relatively complex domains with thousands or millions of states is infeasible for practical application. As Barrett et al. (Barrett, Stone, and Kraus 2011) observed, ad hoc agents can often collaborate effectively when only observing the behavior of an unknown team in a comparatively small section of the domain's state space. This is particularly true in cases where the team attempts repeated trials with static initial conditions. It is, therefore, beneficial for agents to reason over what information about teammates they already possess and evaluate what *subset* of the team's behavior would be beneficial to know.

Decision theory provides a mechanism of determining the expected change in expected utility for a communicative act, as shown in equation 1. The utility of a communicative act is dependent on two factors: the uncertainty regarding which action a collaborator will take and the difference in utility given each potential outcome. During evaluation of such communicative acts, for each potential action response for a state query, the agent reevaluates its expected utility (denoted by V') for both its original policy, π, as well as an updated policy, π', constructed under the new information. The difference in these two utilities is the increase in expected utility the agent would receive given the information *in advance*, allowing the agent to preemptively adapt its plan. The responses are weighted by the predicted probabilities of the actions from the agent's teammate model. Once the set of potential queries has been evaluated, the agent selects the query with the maximal expected change in utility and may repeat this process while queries retain non-zero utility. As our rational agent only alters its policy monotonically when presented with new information, each potential outcome of a communicative exchange results in a non-negative change in expected utility. Likewise, the weighted sum of these potential outcomes—the expectation of change in expected utility—is non-negative. In short, having additional information is never expected to penalize the coordinating agent, though it may result in no change in expected utility.

$$U_{Comm}(s) = \sum_{a \in A} Pr(a|s) \left(V'_{\pi'|a}(s_0) - V'_{\pi|a}(s_0) \right) \quad (1)$$
$$= \left(\sum_{a \in A} Pr(a|s) V'_{\pi'|a}(s_0) \right) - V_\pi(s_0)$$

A complicating factor of the granular state communication problem is the quantity of states to be evaluated. Calculating a utility maximizing policy within a finite horizon using dynamic programming requires time on the order of $h|A||S|^2$. To repeat this process for every potential response in $|A|$ to every possible query state in S then requires time on the order of $h|A|^2|S|^3$. However, we observe that updating the action probability function for a single state, even when found at various horizons in the planning process, leaves many of the previously calculated values unchanged. The change in expected utility need only be propagated from the query state to the origin state of the planner, reevaluating the agent's policy only at states bridging the query state and the origin. We outline this process in Algorithm 1. Once values have been computed for states included within the finite horizon, the evaluation of all potential state-action queries only requires an order of $h|S||A|^2$ time.

Algorithm `EvaluateQuery` (*State s*)
1 For each action $a \in A_s$
2 $v'_s \leftarrow R(s) + \sum_{s'} T(s,a,s')V(s')$
3 $V'_{\pi'|a} \leftarrow$ PropagateValue(s,v'_s)
4 EndFor
5 $U_{Comm} \leftarrow \left(\sum_{a \in A_s} Pr(a|s) V'_{\pi'|a} \right) - V_\pi(s_{origin})$
6 **return** U_{Comm}

Procedure `PropagateValue` (*State s, Value* v'_s)
1 While $s \neq s_{origin}$
2 $s_p \leftarrow predecessor\_state(s)$
3 $v'_{s_p} \leftarrow R(s_p) + \sum_{a \in A} Pr(a|s_p)(T(s_p,a,s)v'_s$
4 $+ \sum_{s' \neq s} T(s_p,a,s')V_\pi(s'))$
5 $s \leftarrow s_p$
6 $v'_s \leftarrow v'_{s_p}$
7 EndWhile
8 **return** v'_s

Algorithm 1: This procedure evaluates a query for a given state by calculating expected changes in the value function, V_π. Here we assume a finite horizon planner has precomputed values V_π. A new value, $V'_{\pi'}$, for an adapted policy, π', is calculated by propagating changes in the value function to the origin state for the planner.

Evaluation

It is of interest to the ad hoc teamwork community to characterize the benefits of intentional communicative acts for collaboration. We test our approach under a varied set of constraints in order to highlight several facets of communication when added to traditional coordinating ad hoc agents.

Domain

The multiagent pursuit problem is a common domain for multiagent systems research and has been adopted for several existing works within the ad hoc teamwork community (Barrett, Stone, and Kraus 2011; Barrett et al. 2013; Sarratt and Jhala 2015). In this domain, a team of agents is tasked with trapping a prey which flees in a toroidal grid. We use a modified version of this concept, similar to (Sarratt and Jhala 2015), in which a team of two agents one of a group of fleeing agents in a maze. In addition to learning how an unknown teammate will pursue a prey, an ad hoc agent must identify which prey is being pursued. We test the communicating ad hoc agent in a maze shown in Figure 1, which

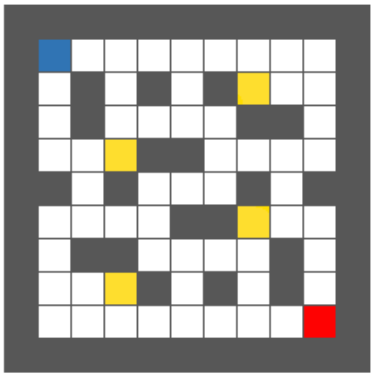

Figure 1: The maze used for the pursuit experiments. The coordinating agents are represented by red and blue cells at the corners of the maze. The fleeing prey are represented by the four yellow cells.

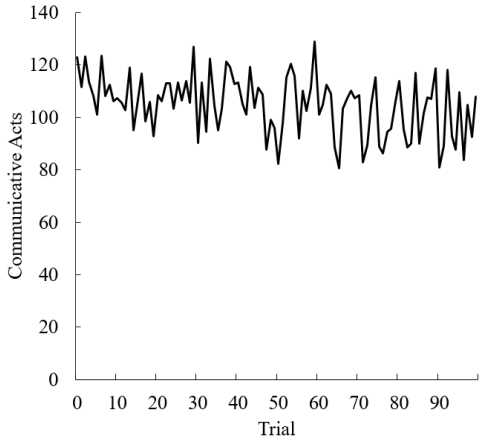

Figure 2: Number of queries by the *No Priors* agent over successive trials.

depicts the initial configuration of the team and the fleeing prey. While deceptively simple, 5.15×10^{10} unique placements of the agents and prey are possible within the maze, with 5.54×10^{7} potential capture states. As such, the domain is large enough to be intractable to solve exhaustively yet small enough for online planning without the necessity of domain-engineered considerations, which may confound the evaluation of our approach.

Agents

We test the communicative capability with two coordinating ad hoc agent types. The first, which we will refer to as the *No Priors* agent, initially possesses a model of its teammate which uniformly predicts the actions of its teammate. This model is updated by observing the teammate, predicting future actions by a frequency count with Laplace smoothing, as shown below.

$$Pr(a|s) = \frac{freq(a|s) + 1}{\sum_{a_i \in A}(freq(a_i|s) + 1)}$$

The second type of coordinating agent, which we will call *With Priors*, utilizes a set of known models to predict teammate actions. Commonly, these models can be learned offline from previous experience or be authored models of simple behavior within the domain (Barrett et al. 2013; Sarratt and Jhala 2015). The agent updates a belief distribution, initially uniform, over the models, $m \in M$, according to Baye's rule using an exponentiated loss function, shown below:

$$Pr_t(m|a) = \frac{Pr(a|m) \times Pr_{t-1}(m)}{\sum_{m_j} Pr(a|m_j) \times Pr_{t-1}(m_j)} \quad (2)$$

with

$$Pr(a|m) \propto e^{-L}, \quad (3)$$

where L is a binary loss function with a value of 1 if the model incorrectly predicts the observed action and 0 otherwise.

Both agents in our experiments plan using Upper Confidence Bounds for Trees (UCT) (Kocsis and Szepesvári 2006), a version of Monte-Carlo tree search (MCTS) balancing exploration of the state space with exploitation of well-performing sampled policies. UCT has been shown to be effective in large POMDPs (Silver and Veness 2010) and is commonly used in similar ad hoc team research (Barrett et al. 2013; Sarratt and Jhala 2015).

The Unknown Teammate In order to ensure a degree of uncertainty in the paired teammate's behavior, we test the coordinating ad hoc agents with a teammate whose behavior is both noisy and inconsistent. At the start of the trial, it will select a target randomly. During 90% of the turns, the teammate will pursue its current target, while it will select a random action with 10% probability. Furthermore, with each step, the teammate may switch targets with a probability given by

$$Pr_{switch} = 0.2 \times \frac{D_{current\,target}}{\sum_{target} D_{target}} \quad (4)$$

where D_{target} is the shortest distance to a given target.

Information Trade-off Over Repeated Trials

Communicating intentional information is proposed to handle cases when an agent is uncertain which action a teammate will take, with potentially large utility differences between the possibilities. There are two main sources of this uncertainty:

1. Inconsistency in behavior - across many observations of a state, a teammate has taken multiple actions many times each.
2. Lack of information in the model - typically this occurs when an agent has not observed a particular state frequently enough to learn a teammate's behavior.

In the former case, the coordinating agent is uncertain which of multiple established teammate strategies. As an example, consider the unfamiliar teammate begins in the lower

right corner of the maze. Across multiple trials, the collaborating agent observes its teammate either proceed north to pursue the prey in the top right corner or proceed west in pursuit of the bottom left prey. After many trials, the coordinating agent expects the teammate to choose either of these two strategies, and depending on the decision-theoretic value, it may query its teammate to determine in which direction it will proceed.

In the latter case, the agent simply does not possess enough information to accurately predict the actions of its teammate. This frequently occurs when initially coordinating with a new teammate or when the system enters into a part of the domain's state space that has not been explored with the teammate and, therefore, lacks observations. In this context, we would expect more communicative acts in unfamiliar territory. Over time, as the agent adjusts its model to fit the teammate's behavior, the agent has less uncertainty regarding the eventual actions it will observe, resulting in diminished communication. This is reflected in the theory of shared mental models (Orasanu 1994), where synced team expectations regarding the status of a task and the individual responsibilities of team members results in lessened conflict and infrequent communication.

We demonstrate this result in our communicative *No Priors* agent over a series of one hundred successive trials with the unfamiliar teammate. At each step, the agent selects all queries with positive utility. Across the trials, the agent retains its model of the teammate's observed behavior. This process is repeated twenty times, and the average number of communicative acts per trial is reported in Figure 2. The data forms a weak negative trend, producing a Spearman correlation coefficient of $\rho = -0.336$ ($p < 0.001$).

Cost-restricted Communication

Typically, a cost, C, is associated with communicating. If two robots are collaborating on a task, they must expend time and energy in order to exchange information. We model this consideration using a fixed utility loss for all communicative acts, though other schemes of assigning costs are possible. When communication is not free, an agent must consider whether the potential utility gain is worth the penalty of transmitting information. Therefore, an agent will only communicate if $U_{Comm} > C$.

Figure 3 illustrates the effect of cost on the communication process over successive queries. For each tested cost of communication, the agent is allowed to query its teammate for policy information as long as each query's utility exceeds the cost of communication. The results are averaged over one hundred trials for each cost. Clearly, increased costs act as a filter over the potential queries an agent may consider. Therefore, in high cost scenarios, agents may only communicate rarely, relying instead on decision theoretic planning under larger uncertainty regarding a teammate's behavior. With lower communicative costs, agents exchange information more readily, allowing for reduced uncertainty and increased expected utility.

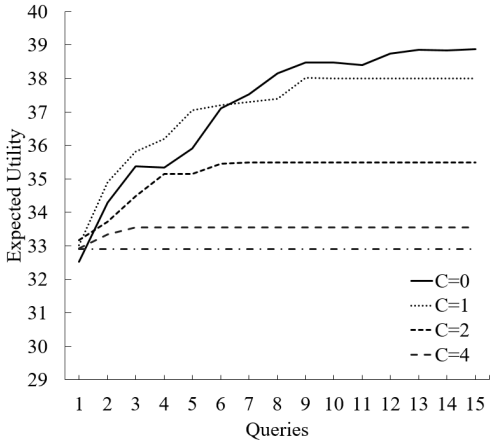

Figure 3: Progression of agent's expected utility over successive queries under various communication costs.

Queried States and Policy Changes

When evaluating a potential state query, three elements factor into the value of the communicative act. First, large differences in utility between actions taken at the state provide greater changes in expected utility when actions are pruned from consideration. In the tested domain, this primarily occurs at the cusp of a capture. Both teammates must enter the prey's cell to capture it. If the teammate switches targets or performs a random action, it may miss the window for capture, allowing the prey to slip by flee into the maze, forcing the team to pursue it until they can surround it once more and attempt capture. This can occur in nearly every location within the maze.

A second consideration in the evaluation of a query is the target state's depth within the planning horizon. As the sequence of actions required to transition to a given state accumulates action probabilities $0 \leq Pr(a|s) \leq 1$ as well as transition probabilities (in stochastic domains) $0 \leq T(s, a, s') \leq 1$, the value of a query is biased toward states closer to the origination of the planning process. Furthermore, as all trials tested begin at the same state but may play out uniquely, we expect common queries across trials earlier, before playouts diverge into unique sections of the state space. This is reflected in Figure 4 which depicts heatmaps of the teammate's location across queries as well as changes in the agent's policy resulting from the communicative acts.

Finally, the uncertainty within a learned model of a teammate is a prominent factor. Consider the progression of the *No Priors* agent. Initially, all action predictions are uniform, providing the maximum uncertainty while planning. Over time, the agent observes consistency in the teammate's behavior within certain states. For example, the teammate rarely doubles back in a hallway. Rather, it maintains momentum in its movement. However, despite potentially numerous observations, branch points may retain their uncertainty to a degree, particularly if the agent has taken each branch with equal frequency. We observe that the local maxima within the queried states (shown as well in Figure 4)

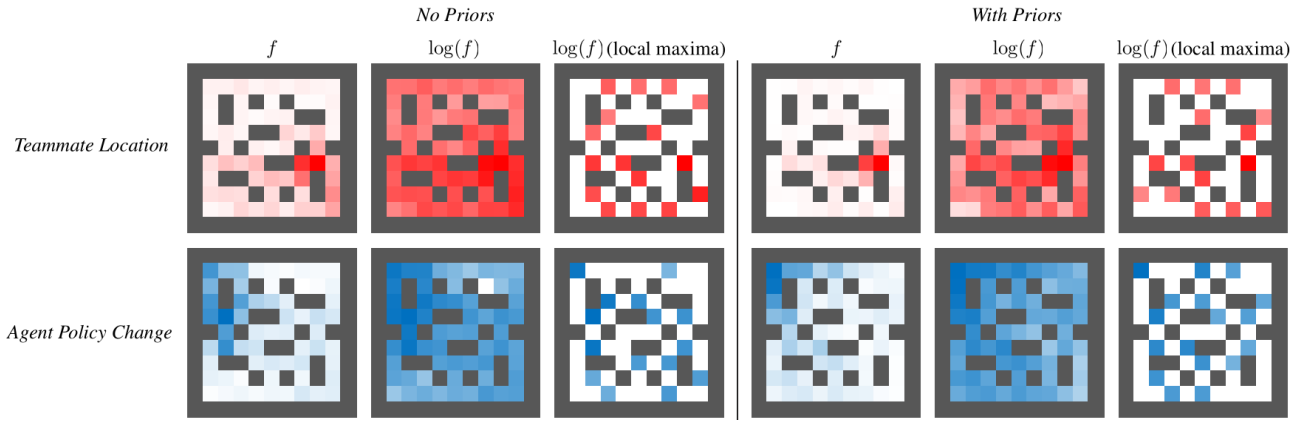

Figure 4: Heatmaps for the queries chosen by the agent when coordinating with an unknown teammate. The first row represents the frequencies, f, of the potential teammate locations in the states queried, while the second row depicts where the agent's policy is changed as a result of the queries. Due to the exponential drop-off in query frequencies radiating out from the initial state, the *log* of the query frequencies is also shown. Finally, all locations except the local maxima are removed in order to identify common, highly valued queries across the state space.

occur primarily at branching points within the maze. Moreover, the local maxima for policy changes also occur at such points, emphasizing the importance of such decision points.

Discussion and Future Work

Communication of intentions in ad hoc team domains is a consideration not to be overlooked. Past work often dismisses the possibility of communication, citing the lack of a shared communication protocol. In domains where communication is permitted, it is commonly added as a small number of high level, domain-specific messages (such as *want to play defense* in RoboCup (Genter, Laue, and Stone 2015)), or it is restricted to sharing only hidden state information (Barrett et al. 2014), as in traditional multiagent systems applications.

When cooperating with unknown teammates, agents with the capability to exchange policy information can act in a proactive manner, acquiring valuable team behavior information early enough that they may adjust their individual plans to further the coordinated effort. In contrast, restricting the agent to learning purely through observation requires that the agent must first observe the act in question or attempt to generalize predictions between models (Barrett et al. 2013) using a small set of related observations. However, the communication of intentional information is not intended as a replacement for traditional learning techniques. Rather, it complements learning agents, as such communicative behavior both requires a reflective analysis of the uncertainty within an existing teammate model and advances the information an agent possesses about its teammates. This motivates further exploration into ad hoc agents with both capacities.

An immediate extension to this work would consider the communication of multiple state-action pairs without independent evaluation. It is possible for two states to have no utility for communication individually but have non-zero utility when considered together. This opens up a combinatorial space of potential intentional information sets that could be communicated, similar to problem of picking a subset of observations to share within a team, as explored by Roth et al. 2006. Due to the intractable nature of the problem, the authors motivated the exploration of heuristics as approximate solutions. Similar techniques to those in (Roth, Simmons, and Veloso 2006) may be beneficial for narrowing the space of the $2^{|S|}$ potential collections of states to be queried.

As a final point of discussion, one characteristic of the decision-theoretic approach to intentional communication should be emphasized: it does not attempt to learn the entire policy of a coordinating teammate. Rather, it evaluates which portions of the policy are worth knowing, that is, which are potentially likely to alter the ad hoc agent's individual plan and improve the agent's utility. In other collaborative planning frameworks, such as SharedPlans (Grosz and Kraus 1999), the hierarchical joint plan is elaborated into complete group and individual plans for the tasks involved. It is conceivable that the members of a team could coordinate effectively without possessing complete knowledge of the individual plans, which is of particular importance when communication is costly. Rather than share the entirety of the joint plan, agents may rely on their predictive capabilities for the subplans of their collaborators. If the agent is uncertain how a teammate will accomplish a task, it can ask a the teammate to elaborate its the entirety or even simply a small section of its plan in a similar fashion as presented in this paper.

References

Agmon, N., and Stone, P. 2012. Leading ad hoc agents in joint action settings with multiple teammates. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 341–348. International Foundation for Autonomous Agents and Multiagent Systems.

Albrecht, S. V., and Ramamoorthy, S. 2012. Comparative evaluation of mal algorithms in a diverse set of ad hoc team problems. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 349–356. International Foundation for Autonomous Agents and Multiagent Systems.

Albrecht, S. V., and Ramamoorthy, S. 2014. On convergence and optimality of best-response learning with policy types in multiagent systems. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 12–21.

Albrecht, S. V.; Crandall, J. W.; and Ramamoorthy, S. 2015. An empirical study on the practical impact of prior beliefs over policy types. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Barrett, S., and Stone, P. 2014. Cooperating with unknown teammates in robot soccer. In *AAAI workshop on multiagent interaction without prior coordination (MIPC 2014)*.

Barrett, S.; Stone, P.; Kraus, S.; and Rosenfeld, A. 2012. Learning teammate models for ad hoc teamwork. In *AAMAS Adaptive Learning Agents (ALA) Workshop*.

Barrett, S.; Stone, P.; Kraus, S.; and Rosenfeld, A. 2013. Teamwork with limited knowledge of teammates. In *AAAI*.

Barrett, S.; Agmon, N.; Hazon, N.; Kraus, S.; and Stone, P. 2014. Communicating with unknown teammates. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 1433–1434. International Foundation for Autonomous Agents and Multiagent Systems.

Barrett, S.; Stone, P.; and Kraus, S. 2011. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 567–574. International Foundation for Autonomous Agents and Multiagent Systems.

Chakraborty, D., and Stone, P. 2013. Cooperating with a markovian ad hoc teammate. In *Proceedings of the 2013 international conference on Autonomous agents and multiagent systems*, 1085–1092. International Foundation for Autonomous Agents and Multiagent Systems.

Genter, K.; Laue, T.; and Stone, P. 2015. The robocup 2014 spl drop-in player competition: Encouraging teamwork without pre-coordination. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1745–1746. International Foundation for Autonomous Agents and Multiagent Systems.

Gmytrasiewicz, P. J.; Durfee, E. H.; and Rosenschein, J. 1995. Toward rational communicative behavior. In *AAAI Fall Symposium on Embodied Language*, 35–43.

Gmytrasiewicz, P. J.; Durfee, E. H.; and Wehe, D. K. 1991. The utility of communication in coordinating intelligent agents. In *AAAI*, 166–172.

Grosz, B. J., and Kraus, S. 1999. The evolution of sharedplans. In *Foundations of rational agency*. Springer. 227–262.

Kocsis, L., and Szepesvári, C. 2006. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*. Springer. 282–293.

Orasanu, J. 1994. Shared problem models and flight crew performance. *Aviation psychology in practice* 255–285.

Pynadath, D. V., and Tambe, M. 2002. Multiagent teamwork: Analyzing the optimality and complexity of key theories and models. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, 873–880. ACM.

Roth, M.; Simmons, R.; and Veloso, M. 2006. What to communicate? execution-time decision in multi-agent pomdps. In *Distributed Autonomous Robotic Systems 7*. Springer. 177–186.

Roth, M.; Simmons, R.; and Veloso, M. 2007. Exploiting factored representations for decentralized execution in multiagent teams. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 72. ACM.

Sarratt, T., and Jhala, A. 2015. Tuning belief revision for coordination with inconsistent teammates. In *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*.

Silver, D., and Veness, J. 2010. Monte-carlo planning in large pomdps. In *Advances in Neural Information Processing Systems*, 2164–2172.

Stone, P.; Kaminka, G. A.; Kraus, S.; Rosenschein, J. S.; et al. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI*.

Stone, P.; Kaminka, G. A.; Kraus, S.; Rosenschein, J. S.; and Agmon, N. 2013. Teaching and leading an ad hoc teammate: Collaboration without pre-coordination. *Artificial Intelligence* 203:35–65.

Stone, P.; Kaminka, G. A.; and Rosenschein, J. S. 2010. Leading a best-response teammate in an ad hoc team. In *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets*. Springer. 132–146.

Tambe, M. 1997. Agent architectures for flexible. In *Proc. of the 14th National Conf. on AI, USA: AAAI press*, 22–28.