UNIVERSITY OF CALIFORNIA
SANTA CRUZ

COMMUNICATING PLANS IN AD HOC MULTIAGENT TEAMS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Trevor Santarra

March 2019

The Dissertation of Trevor Santarra
is approved:

Professor Arnav Jhala, Chair

Professor Pieter Spronck

Professor Adam Smith

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Communicating Plans in Ad Hoc Multiagent Teams

by

Trevor Santarra

With the rising use of autonomous agents within robotic and software settings, agents may be required to cooperate in teams while having little or no information regarding the capabilities of their teammates. In these ad hoc settings, teams must collaborate on the fly, having no prior opportunity for coordination. Prior research in this area commonly either assumes that communication between agents is impossible given their heterogeneous design or has left communication as an open problem. Typically, to accurately predict a teammate's behavior at a future point in time, ad hoc agents leverage models learned from past experience and attempt to infer a teammate's intended strategy through observing its current course of action. However, these approaches can fail to arrive at accurate policy predictions, leaving the coordinating agent uncertain and unable to adapt to its teammates' plans. We introduce the problem of communicating minimal sets of teammate policies in order to provide information for collaboration in such ad hoc environments. We demonstrate how an agent may determine what information it should solicit from its peers but further illustrate how optimal solutions to such a problem have intractable computational requirements. Nonetheless, through the characterization of this difficulty, we identify strategies that permit approximate or heuristic approaches, allowing the practical application of this capacity in ad hoc teams.

To Bethany, Julie, and Eric, for inspiring me to hope for a future where artificial intelligence can work *with* us to improve our lives and our world.

Acknowledgments

I would like to thank my adviser, Arnav Jhala, for always possessing the willingness to dive down a new research direction with me, no matter how unfamiliar it is. I am grateful to have had a mentor who trusted my intuition and defended my ideas at every opportunity. I would not be the researcher I am today had it not been for Arnav.

Additionally, I would like to thank the remaining members of my committee, Pieter Spronck and Adam Smith, for their excitement, feedback, and kind wishes.

I will not forget the many long hours spent with Michael Leece and Morteza Behrooz, whether discussing papers, sharing a drink, or bonding over a board game.

I would also like to extend my thanks to Nicolas Meuleau, who brought me into a small team with a grand vision. To love my work has been a luxury, and were it not for such an opportunity, I may never have renewed my confidence in my own ability to finish this research.

Lastly, I am indebted for the resources provided by the Open Science Grid [105, 126], which is supported by the National Science Foundation award 1148698, and the U.S. Department of Energy's Office of Science. The experiments of Chapter 11 would not have been accomplished without the substantial computational resources available through the Open Science Grid.

Chapter 1

Introduction

In traditional multiagent systems (MAS) literature, teams of agents share an identical design for reasoning, planning, and executing actions, allowing perfect modeling of their teammates. Ad hoc teamwork [134] introduces the notion that with the increasing utilization of autonomous agents and robots, agents of heterogeneous design may need to coordinate in real world settings. The authors cite the ability of humans work together without prior coordination or prior knowledge of one another as a capacity to be pursued by the multiagent systems community. In these scenarios, teammates may have little to no experience cooperating with their peers, putting the collaborators in a position where they must learn the capabilities of their teammates during the act of coordination. As such, the task of ad hoc teamwork can be conceptualized as the interplay of many concepts: reasoning under uncertainty, agent modeling, plan recognition, machine learning, and—as the primary interest of this work—communication.

Much of the existing ad hoc teamwork literature focuses on reinforcement learning and decision-theoretic planning as the primary mechanisms for tackling this challenge. Agents use models of known behavior to predict an ad hoc agent's actions, employing probabilistic reasoning to attempt coordination in instances

where the strategies of their teammates are uncertain [12, 14, 10, 11, 3, 5, 6, 7, 4, 31, 33]. These models can be refined, combined, or learned anew during execution by observing teammate behavior, increasing the accuracy of the predictions over time, which results in more consistent coordination. However, it is often the case that despite leveraging both prior experience and inference from observed behavior, an agent may be left uncertain as to intended actions of its teammates. This is particularly true when coordinating with a teammate whose policy—the conditional plan an agent follows—stands apart from any behavior previously experienced by the agent.

Ad hoc teamwork possesses many similarities to other multiagent coordination problems, particularly those in partially observable domains where agents must make inferences regarding the state of the world they cannot observe directly. In our case, agents cannot know the full plans of their collaborators. In such information asymmetric domains, it is often necessary to share information between agents, effectively syncing the team's beliefs about the world and, as a result, the individual policies of each member. This exchange of information can be a critical aspect of coordination.

Communication within ad hoc teams has been an open problem within the community since its inception. In many instances, where agents are designed by different institutions or are built with unique technologies, communication between teammates may be impossible, each possessing incompatible forms of information broadcasting and receiving. Despite this notion, many domains exist in which heterogeneous agents coordinate while possessing the capacity for communication. One compelling application is that of human-agent teams, whether with virtual or physically embodied agents, where natural language or specialized interfaces facilitate communication. What follows is a natural question: *What*

should members of an ad hoc team communicate?

Consider a scenario in which two rescue robots must search an environment for trapped persons and then coordinate to remove debris and deliver supplies. Naturally, the agents may need to split up to cover more ground. But what areas should each agent search independently? If a trapped person is found, when and how should the agent attempt to rendezvous with its teammate? Finally, once the team has been notified, who should take on the responsibility of moving debris or fetching needed supplies? These questions demonstrate the interdependence of the individual agents' plans. One robot can only be confident in its own actions with some degree of certainty regarding the intentions of its teammates.

If one primary use of communication is the reduction of uncertainty in partially observable domains, we propose that uncertain policies of teammates serve as a natural target for explicit communication. In this regard, we treat teammates as oracles for their own individual policies, having perfect knowledge of or at least being capable of computing answers to queries regarding their policies. It is left to an ad hoc agent, then, to decide what information it requires from its teammates in order to improve the predictions of its teammates' plans and adapt its own.

In contrast to traditional multiagent communication applications where communicative acts are few in number, agent reasoning over uncertain teammate policies may consider the entire space of decisions a teammate may make throughout the collaborative effort. Furthermore, in choosing a sequence of decision points to clarify, an agent must consider how the potential responses it may receive influence which policy piece of information it queries next. In this thesis, we formalize this problem as a Markov decision process (MDP), providing a direct mechanism for computing these conditioned communication policies but at the expense of covering a state space exponentially larger than the original coordination domain.

This exponential increase in complexity motivates the exploration of approximate or heuristic approaches, one of the primary focuses of this work. Accordingly, the primary contributions of this thesis are

- A formal characterization of the decision problem of identifying elements of teammate policies which should be explicitly communicated,
- A theoretical analysis of the problem, illustrating its complexity as well as various characteristics which may be informative for developing practical solution methods,
- An approximate, heuristic-based, decision-theoretic planning approach for computing communication policies, and
- An empirical evaluation of the proposed technique in an ad hoc multiagent teamwork domain.

This dissertation is organized as follows:

- **Chapter 2** outlines the construction of the perspective of an ad hoc agent operating within an multiagent team while uncertain of its teammates' individual policies.
- **Chapter 3** summarizes related work on coordination in ad hoc teams, with a particular focus on approaches for managing beliefs, learning models of teammate behavior, and sharing information through explicit communication. We situate the techniques evaluated in this thesis among the such literature.
- **Chapter 4** introduces the multiagent pursuit domain, of which we use a two-agent version for empirical evaluation of approaches throughout the document.

- **Chapter 5** covers a non-communicative example of ad hoc teamwork. We propose a belief revision approach over known models, under which a coordinating agent can detect and adapt to changes in teammate behavior.

- **Chapter 6** motivates the need for acquiring information beyond simply observing the behavior of other agents. For this purpose, we propose explicit communication.

- **Chapter 7** formally defines the problem of communication in ad hoc teams as an extension to the perspective outlined in Chapter 2.

- **Chapter 8** characterizes the communication problem, providing both theoretical and qualitative analysis, most notably the substantial computational needs for determining communication policies in practice.

- **Chapter 9** attempts a greedy approach to the communication problem in the two-agent pursuit domain.

- **Chapter 10** proposes a set of measures to provide a heuristic ordering of queries. We leverage this rough ordering to prune the space of potential queries an agent considers while computing a communication policy.

- **Chapter 11** evaluates the proposed heuristics and assesses the effects of communication cost, agent experience, and domain choice on the success of coordinating agents.

- **Chapter 12** provides a retrospective on the work, discussing its initial successes in establishing a practical approach to communicating in ad hoc teams. Furthermore, we discuss a myriad of extensions and new directions of the work.

Chapter 2

Background and Notation

Multiagent systems (MAS) is a field concerned with the interaction of intelligent agents within an environment. In an interactive space, the actions of fellow agents in pursuit of their individual goals may directly affect the shared world state in which an agent is planning. For this reason, an agent is forced to consider the actions of other agents as it plans its own course of actions, whereas in spaces in which the actions of agents do not alter any shared resources, an agent may adopt a single-agent perspective, effectively ignoring other agents.

The sizes of the domain state space, the potential set of potential interactions between agents, and the simultaneous actions possible often characterizes such multiagent problems as intractable to consider in their entirety, motivating the division of labor across the agents. The successful resulting decentralized control of a process is dependent on the capability of each individual agent to process local information, reason about the actions of other agents, and adapt its plan accordingly. If there is uncertainty regarding the potential behavior of another agent, the planning process incurs the added difficulty of recursively modeling the potential utility function, goals, beliefs, and other factors that inform an agent's decisions [54, 52, 21, 108, 42].

One of the foremost hurdles for multiagent team decision problems is computational complexity required to compute a agent's policy. Many MAS domains can be represented by decentralized control of Markov decision processes, called DEC-MDPs and DEC-POMDPs for the fully and partially observable versions, respectively. The complexity of finding an optimal joint policy in these representations is known to be NEXP-complete in the finite horizon case and undecidable in the infinite horizon case [21].

Fortunately, it is often possible to reduce the complexity to a degree by allowing simplifying assumptions and accepting locally optimal solutions. For example, Nair et al. [91] propose fixing teammate policies and searching for locally optimal agent policies until an equilibrium is reached, resulting in a significant reduction in computational time. In the vein of simplifying the problem directly, agents who are designed with or gain perfect information regarding the state and decision process of other agents can then act as though controlled by a centralized process [100]. For example, for DEC-POMDPs, providing an agent with other agents' action and observation histories—either via the domain itself or through free communication between agents—allows the reduction of the decision problem to an equivalent single-agent POMDP [109]. POMDP solvers have had considerably more advances than their decentralized counterparts and are frequently solved via dynamic programming [16] or sample-based techniques [129].

Within domains where such synchronization across agents is not possible, Tambe [140] and Gmytrasiewicz and Durfee [52] stress the importance of recursive agent modeling, estimating other agents' beliefs over attributes of the world, their beliefs regarding each other agent, as well as their beliefs over other agents' beliefs and so on. Potentially, such recursive modeling can be infinitely deep, though the assumption of bounded rationality [130] can often yield more human-

like results [107]. Moreover, Mundhe and Sen [89] observed that modeling other agents as being non-recursive and having fixed, probabilistic policies can still lead to convergence to optimal policies in certain domains.

In this chapter, we construct the decision problem faced by an agent coordinating in an ad hoc multiagent team. We begin with the model of a multiagent MDP [25], providing the framework for a team decision problem. By extending the model with uncertainty over the underlying policies of an agent's teammates, we illustrate how ad hoc agents define and revise beliefs regarding their collaborators' plans.

2.1 Multiagent Markov Decision Processes

The control of a process by multiple coordinating agents is represented as a Multiagent Markov Decision Process (MMDP) [25], which can be thought of as a generalization of single-agent Markov Decision Processes (MDPs). Formally, an MMDP for a team of $n+1$ agents is defined by the tuple $\langle S, \bar{A}, T, R, \gamma \rangle$ where:

- S is a finite set of states, fully observable by all agents;
- $\bar{A} = \bigtimes_{i=0}^{n} A_i$ is the set of joint actions given by the Cartesian product over the sets of individual agent actions, A_i;
- $T : S \times \bar{A} \times S \longmapsto [0, 1]$ is a transition function which provides the likelihood of a state transition given a joint action;
- $R : S \times \bar{A} \times S \longmapsto \mathbb{R}$ is a real-valued reward function;
- $\gamma \in [0, 1)$ is a discount factor.

MMDPs provide a computational model for computing an optimal joint policy, $\bar{\pi} : S \longmapsto \bar{A}$, maximizing the expected cumulative reward over a finite or infinite

horizon, $\mathbb{E}\left[\sum_t \gamma^t R_t \middle| \bar{\pi}\right]$. The computation of an an optimal MMDP policy can be performed either via a centralized process, which computes all of the agents' individual policies, or in a decentralized fashion, in which each agent computes its own policy as part of the greater joint policy. For now, we will discuss the former, leaving the challenges of the decentralized version for Section 2.2.

For a centralized process, we formulate recursive value functions for the MMDP in the same manner as typically used for single-agent MDPs. The recursive Bellman equation for the MMDP acting under joint policy, $\bar{\pi}$, is given by

$$V_{\bar{\pi}}(s) = \sum_{s' \in S} T(s, \bar{\pi}(s), s') \Big[R(s, \bar{\pi}(s), s') + \gamma V_{\bar{\pi}}(s') \Big]$$

A policy is considered optimal if and only if $V_{\bar{\pi}}(s) \geq V_{\bar{\pi}'}(s)\ \forall s \in S,\ \forall \bar{\pi}'$. It is important to note that for a given MMDP, there may exist multiple optimal joint policies, each of which induces the optimal value function $V^*(s)$ given by

$$V^*(s) = \max_{\bar{a} \in \bar{A}} \sum_{s' \in S} T(s, \bar{a}, s') \Big[R(s, \bar{a}, s') + \gamma V^*(s') \Big]$$

.

2.2 Challenges of Decentralized Teamwork

It is often the case that agents coordinating on a joint problem cannot utilize a centralized process to compute and assign the individual policies that comprise an optimal joint policy. In these situations, it is left up to each agent to decide on an individual policy that coordinates successfully as a part of the larger joint policy. However, this is a non-trivial step, as it can require the entire team of agents to arrive at compatible policies. Consider the case where a clear joint plan is established, but that the overall structure is divided into relatively independent

roles for each of the agents fill. How are the roles assigned? If, due to some inherent flaw in the decentralized assignment of roles, two agents attempt to fill the same role, the joint plan may fail. This hurdle is compounded when multiple optimal joint policies exist, as each agent must additionally identify which joint policy to pursue before attempting to assign an individual policy. Furthermore, in situations where the scale of the problem necessitates approximate solutions, the space of potentially joint policies can grow dramatically.

As such, the issue of coordination in multiagent teams has been at the center of much of the extensive work within the multiagent systems community. Many techniques have been proposed, such as locker room agreements for role assignment [137], explicit communication of observations in partially-observable domains [108, 116], iterative advancement toward stable behaviors via machine learning [32, 36], and even negotiating joint plans directly [57, 58]. As described in [25], such techniques largely fall into three categories: imposed conventions, learned coordination, and explicit communication.

Conventions, in this context, refer to established patterns of behavior, often included by the designer of the system, in order to restrict the behavior of the agents in order to achieve coordination. For example, in the event that multiple optimal policies can be achieved, a lexicographic ordering could be imposed and shared across all agents, under which each agent could arrive at consensus without the requirement of explicit coordination acts. For the purposes of this work's focus on ad hoc teams, conventions are removed from consideration, as they are in opposition to the assumption that each agent possesses minimal prior knowledge regarding the decision processes of their teammates.

Learned coordination, however, is the most utilized approach for ad hoc team domains. Given the inherent uncertainty regarding the intentions of new team-

mates, it is useful to observe and identify patterns within the behaviors of the team, in order to adapt accordingly. Much of the existing work focuses on attempting to learn an accurate model of teammate behavior on-the-fly during the act of coordination [13] or to generalize existing models of similar teammates to construct a new, accurate model with only sparse observations [12]. We will return to these approaches in more detail in Chapter 3. Furthermore, while ad hoc teamwork has historically focused on the ability of a single agent to learn to coordinate within a new team, research from other areas of multiagent systems as well as game theory have studied learned coordination from a group perspective, allowing multiple agents to learn and adapt independently such that the team settles into policies that represent Nash equilibria for the joint effort [36].

Finally, communication provides agents the capability to send and receive information necessary for coordination. Often, the particular form of communicative acts relies on the domain-specific application, potentially containing observations, commands, queries, physical signals, or messages encoding relevant information regarding the world [51, 108, 116]. After a communicative act, one or more of the agents may adapt its behavior, improving the overall likelihood of success. Within many ad hoc teamwork domains, communication is assumed to be infeasible, due to the costs involved or to incompatibilities in agents' communication protocols. This is not strictly true, as in the primary example of human ad hoc teamwork—a pickup game of soccer—humans often attempt some form of gestures or other signaling. The existence of even limited forms of communication can provide a useful mechanism for coordination.

2.3 Individual Agents in a Multiagent Team

From the perspective of a single agent within a decentralized multiagent team, the problem of selecting an individual policy depends directly upon the individual policies of its teammates. We will first examine how an agent decides is own policy when it knows the policies of its teammates. The construction of this perspective introduces this fixed knowledge by extending the state space, such that this fact of the world is known to the agent. Let A_i^S be the space of mappings, $S \mapsto A_i$, and $\mu_i \in A_i^S$ $(i = 0, 1, 2, \ldots n)$ be the deterministic policy implemented by player i. We can construct the single agent's decision problem for agent $i = 0$ as an MDP, $\langle \tilde{S}, A_0, \tilde{T}, \tilde{R}, \gamma \rangle$, such that

- $\tilde{S} = S \times \left(\bigtimes_{i=1}^{n} A_i^S \right)$ is the set of states, extended by possible assignments of policies of the other agents;
- A_0 is the set of actions for the individual agent $(i = 0)$;
- The modified transition function is given by

$$\begin{aligned} \tilde{T}(\tilde{s}, a_0, \tilde{s}') &= \tilde{T}\left(\left(s, \bigtimes_{i=1}^{n} \mu_i \right), a_0, \left(s', \bigtimes_{i=1}^{n} \mu_i' \right) \right) \\ &= \begin{cases} T\left(s, \left(a_0, \bigtimes_{i=1}^{n} \mu_i(s) \right), s' \right) & \text{if } \mu_i = \mu_i' \ \forall i \\ 0 & \text{otherwise}^1 \end{cases} \end{aligned}$$

[1] We assume the policies of other agents do not change, and therefore assign $\tilde{T}(\cdot, \cdot, \cdot) = 0$ when $\mu_i \neq \mu_i'$.

- The modified reward function:

$$\begin{aligned}\tilde{R}(\tilde{s}, a_0, \tilde{s}') &= \tilde{R}\left(\left(s, \bigtimes_{i=1}^{n} \mu_i\right), a_0, \left(s', \bigtimes_{i=1}^{n} \mu_i'\right)\right) \\ &= \begin{cases} R\left(s, \left(a_0, \bigtimes_{i=1}^{n} \mu_i(s)\right), s'\right) & \text{if } \mu_i = \mu_i' \ \forall i \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

As before, the optimal value function is given by

$$\begin{aligned}\tilde{V}^*(\tilde{s}) &= \tilde{V}^*\left(\left(s, \bigtimes_{i=1}^{n} \mu_i\right)\right) \\ &= \max_{a_0 \in A}\left[\sum_{\tilde{s}' \in \tilde{S}} \tilde{T}(\tilde{s}, a_0, \tilde{s}')\left(\tilde{R}(\tilde{s}, a_0, \tilde{s}') + \gamma \tilde{V}^*(\tilde{s}')\right)\right] \\ &= \max_{a_0 \in A}\left[\sum_{s' \in S} T\left(s, \bar{a}, s'\right)\left(R\left(s, \bar{a}, s'\right) + \gamma \tilde{V}^*\left(\left(s', \bigtimes_{i=1}^{n} \mu_i\right)\right)\right)\right]\end{aligned}$$

where $\bar{a} = \langle a_0, \mu_1(s), \ldots, \mu_n(s)\rangle$.

However, in practice, an agent will not have full knowledge of the team's policies. Within the context of ad hoc teamwork, it is necessary to consider how such uncertainty plays a role in the selection of individual policies as well as how learning and communication can aid in coordination. Agent 0 cannot observe the policies $(\mu_1, \ldots, \mu_n)$ of other agents. Instead, it must infer the individual policies through observing the actions taken by other agents. In order to account for such observation-based inference, we reformulate the perspective of the single agent within an ad hoc multiagent team as a partially observable Markov Decision Process (POMDP) [9]. POMDPs provide an establish framework for decision making under uncertainty when aspects of the world cannot be directly observed. In the ad hoc teamwork scenario, the underlying policies of the coordinating teammates are only partially observable—that is, the agent must infer each policy through a

stream of state-action observations which only reveal a subset of the entire policy. This formulation of the ad hoc teamwork problem is common across much of the existing work [7, 11, 14, 122] and demonstrates the heavy computation requirements of optimal decision making in such scenarios. Finding an optimal policy in a POMDP has been shown to be PSPACE-complete [101] and has motivated much work in approximate techniques and domain-specific optimizations. The complexity arises from the uncertainty of the true underlying state. We introduce such partial-observability by extending the single agent's perspective as follows:

- From the previous model we retain the state space $\tilde{S}$, action space A_0, transition function $\tilde{T}$, reward function $\tilde{R}$, and discount factor γ.

- In a given state, agent 0 observes the world state as well as the actions taken by all other agents, as specified by the observation space, $\mathsf{O} = S \times \left(\bigtimes_{i=0}^{n} A_i \right)$. An observation is denoted $\mathsf{o} = \left(s^o, \bigtimes_{i=1}^{n} a_i^o \right)$.

- Furthermore, the observation probability function, Ω, is given by:

$$\begin{aligned} \Omega(\mathsf{o} \mid \tilde{s}, a_0, \tilde{s}') &= \Omega\left(\left(s^o, \bigtimes_{i=1}^{n} a_i^o \right) \;\middle|\; \left(s, \bigtimes_{i=1}^{n} \mu_i \right), a_0, \left(s', \bigtimes_{i=1}^{n} \mu_i \right) \right) \\ &= \begin{cases} \prod_{i=1}^{n} \Pr(a_i^o \mid s, \mu_i) & \text{if } s^o = s' \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

For the purpose of reasoning over the team's individual policies, the coordinating agent maintains beliefs, $b_i : A_i^S \mapsto [0, 1]$, each of which is a probability distribution over A_i^S. Observations, if sufficiently informative, can help diminish this uncertainty significantly. When observing agent i execute action $a_i \in A_i$ in state $s \in S$, the agent revises its belief, $b_i' = \mathcal{B}(b_i, s, a_i)$, where $\mathcal{B}$ is a belief update function. Here, we assume this update applies Bayes' theorem to achieve the

posterior belief:

$$
\begin{aligned}
b_i'(\mu_i) = \mathcal{B}(b_i, s, a_i) &= \Pr(\mu_i | a_i, s) \\
&= \frac{\Pr(a_i | s, \mu_i) \Pr(s, \mu_i)}{\Pr(a_i)} \\
&= \frac{\Pr(a_i | s, \mu_i) b_i(\mu_i)}{\Pr(a_i)} \\
\text{where } \Pr(a_i | s, \mu_i) &= \begin{cases} 1 & \mu_i(s) = a_i \\ 0 & \text{otherwise} \end{cases} \\
\text{and } \Pr(a_i) &= \sum_{\mu_i \in A_i^S} \Pr(a_i | s, \mu_i) b_i(\mu_i)
\end{aligned}
$$

As is common for POMDPs [73], we formulate this as an equivalent *belief* MDP, where each state represents a particular state of of the agent's beliefs, given a sequence of observations and corresponding revision of prior beliefs. For this belief MDP,

- The belief state space $\mathsf{S_b} : \tilde{S} \longmapsto [0, 1]$ is the belief space over $\tilde{S}$, covering both the world state space, S, and the teammate policy space, $A_i^S \ \forall i$. Because the state component of $s \in S$ is fully observable, the belief state space can be denoted

  $$\mathsf{S_b} = S \times \left(\bigtimes_{i=1}^{n} B_i \right)$$

  where $B_i = \mathcal{P}(A_i^S)$ is the set of probability distributions over A_i^S. A generic state is denoted $\mathsf{b} = \left(s, \bigtimes_{i=1}^{n} b_i \right)$, where $s \in S$ and $b_i \in B_i$.

- The belief transition probabilities are given by the formula

$$\begin{aligned}
\mathsf{T}(\mathsf{b}, a_0, \mathsf{b}') &= \sum_{\mathsf{o} \in \mathsf{O}} \Pr\left(\mathsf{b}' \;\middle|\; \mathsf{b}, a_0, \mathsf{o}\right) \Pr\left(\mathsf{o} \;\middle|\; \mathsf{b}, a_0\right) \\
\text{where } \Pr\left(\mathsf{b}' \;\middle|\; \mathsf{b}, a_0, \mathsf{o}\right) &= \Pr\left(\left(s', \bigtimes_{i=1}^{n} b_i'\right) \;\middle|\; \left(s, \bigtimes_{i=1}^{n} b_i\right), a_0, \left(s^o, \bigtimes_{i=1}^{n} a_i^o\right)\right) \\
&= \begin{cases} 1 & \text{if } \mathcal{B}(b_i, s, a_i) = b_i' \; \forall i \text{ and } s' = s^o \\ 0 & \text{otherwise} \end{cases} \\
\text{and } \Pr\left(\mathsf{o} \;\middle|\; \mathsf{b}, a_0\right) &= \Pr\left(\left(s^o, \bigtimes_{i=1}^{n} a_i^o\right) \;\middle|\; \left(s, \bigtimes_{i=1}^{n} b_i\right), a_0\right) \\
&= \Pr\left(s^o \;\middle|\; s, \bigtimes_{i=1}^{n} a_i^o, a_0\right) \Pr\left(\bigtimes_{i=1}^{n} a_i^o \;\middle|\; s, \bigtimes_{i=1}^{n} b_i\right) \\
&= T\left(s, \left(a_0, \bigtimes_{i=1}^{n} a_i^o\right), s^o\right) \prod_{i=1}^{n} \Pr(a_i \;|\; s, b_i)
\end{aligned}$$

- The reward can similarly be given by

$$\begin{aligned}
\mathsf{R}(\mathsf{b}, a_0, \mathsf{b}') &= \mathsf{R}\left(\left(s, \bigtimes_{i=1}^{n} b_i\right), a_0, \left(s', \bigtimes_{i=1}^{n} b_i'\right)\right) \\
&= R\left(s, \left(a_0, \bigtimes_{i=1}^{n} a_i^*(s, b_i')^\dagger\right), s'\right)
\end{aligned}$$

In contrast to the single agent in a multiagent MDP representation, the agent's policy is no longer defined over the state space S; rather, it is a mapping of actions to beliefs, $\pi : B \mapsto A$, and the agent's policy is computed by maximizing the value

†Given a state, s, and the posterior belief, b_i', we can deduce the action performed by player i in s. By construction, b_i' gives 0 probability to every policy μ_i such that $\mu_i(s) \neq a_i$. Let $a_i^*(s, b_i')$ denote the unique action performed by agent i which has positive probability in s according to b_i'.

function,

$$V_\pi(\mathsf{b}) = \sum_{\mathsf{b}' \in \mathsf{S_b}} \mathsf{T}(\mathsf{b}, \pi(\mathsf{b}), \mathsf{b}') \left(\mathsf{R}(\mathsf{b}, \pi(\mathsf{b}), \mathsf{b}') + \gamma V_\pi(\mathsf{b}')\right). \tag{2.1}$$

It is worth noting that the potential size of this belief space can be large, perhaps intractably so, depending on the space of expected policies and the number of teammates being modeled. A single teammate, for example, selects a policy from $|A_i|^{|S|}$ possible policies. If the coordinating agent is modeling n teammates independently, it may—in the worst case—consider $\prod_{i=1}^{n} |A_i|^{|S|}$ potential combinations of individual teammate policies. While much existing research has considered large state spaces, relatively little work has considered ad hoc coordination within large scale multiagent teams. A typical instance of ad hoc teamwork considers an agent coordinating with a single teammate or within a team of agents that have a shared joint policy, allowing a coordinating agent to use observations of individual agent actions to infer the policies across all of its peers [13, 12]. Larger teams composed of many ad hoc agents remains an open topic for research, and we encourage the reader to review results of the Robocup drop-in league for a view of current initiatives in that area [47, 82].

2.4 Planning and Acting in Ad Hoc Teams

In this chapter, we have laid out the perspective of an agent coordinating in an ad hoc team, providing a framework for managing beliefs over the policies of the other team members and computing a coordination policy under such uncertainty. Figure 2.1 depicts the process by which the agent incorporates observations of teammate behavior into its beliefs, which in turn affects its individual policy.

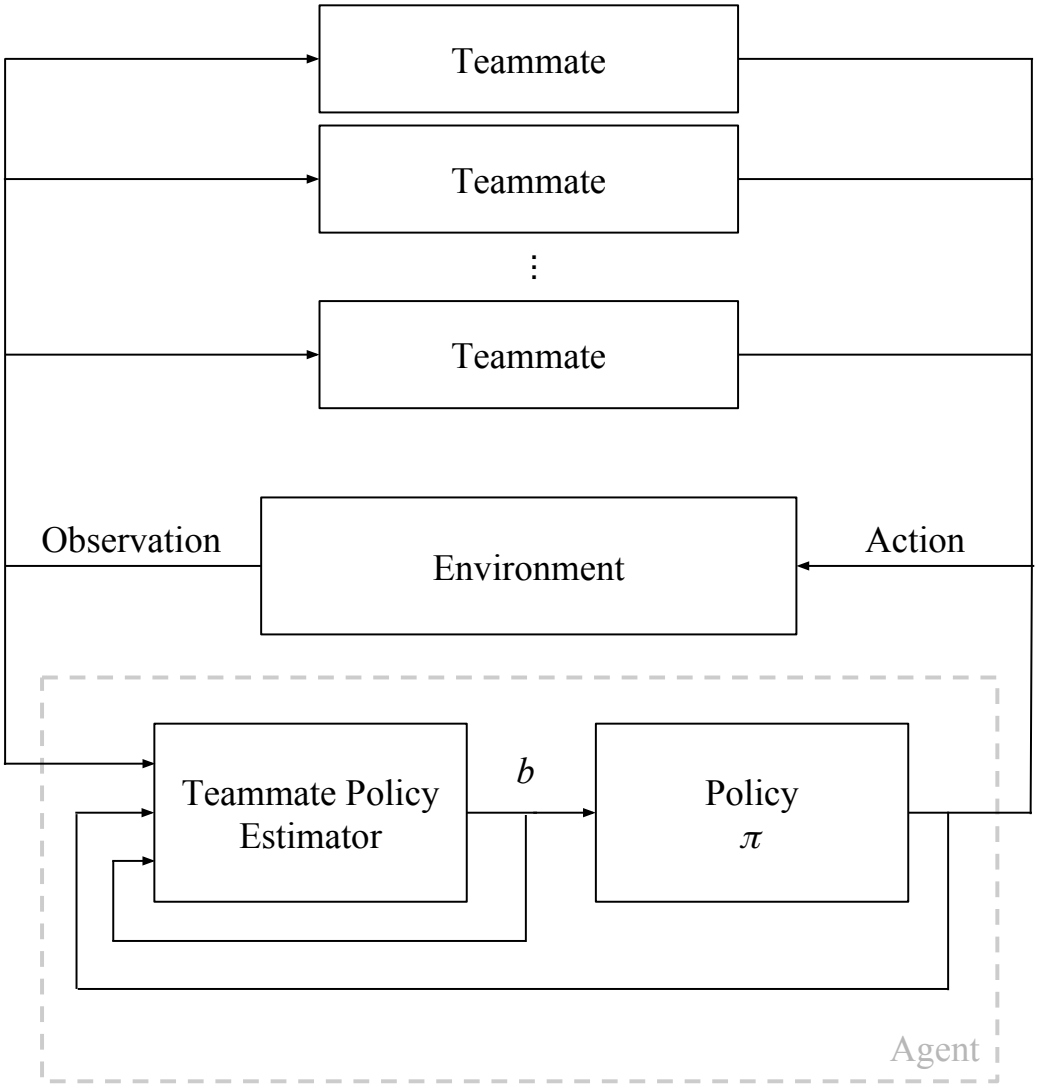

Figure 2.1: A diagram of ad hoc teamwork, in which a team of agents coordinates to mutually affect the environment.

Chapter 3

Related Work in Ad Hoc Teamwork

Stone et al. [134] introduced the challenge of autonomous ad hoc teamwork, citing the ability of humans work together without prior coordination or prior knowledge of one another as a capacity to be pursued by the multiagent systems community. Similar to the human example, ad hoc autonomous agent teams are composed of two or more agents designed separately and with minimal shared information. Specifically, the community was posed the following challenge:

> To create an autonomous agent that is able to efficiently and robustly collaborate with previously unknown teammates on tasks to which they are all individually capable of contributing as team members.

The task of as hoc teamwork can be conceptualized as the interplay of many concepts: multiagent planning with heterogeneous agents, agent modeling, plan recognition, reinforcement learning, and communication. In order to coordinate effectively, an agent must consider the uncertainty of its teammates' behavior, construct a predictive model from observations, communicate asymmetric information, and be able to adapt its own policy accordingly. To illustrate these broad

capabilities, we survey techniques from existing research on ad hoc teamwork.

3.1 Managing Uncertainty

With the minimal assumptions of what constitutes ad hoc teamwork, it is a natural consequence that the existing literature on the topic varies on many fronts, including target application domain, setup of initial beliefs regarding other agents, as well as techniques of managing the uncertainty throughout the collaborative effort.

Consider, first, the initial beliefs an agent holds with regard to its teammates' intended plans. In an ideal scenario, the prior belief would perfectly match the distribution of teammate policies according to the relative likelihoods across the true underlying population from which teammates are drawn. However, this requires either possessing such information a priori or having sampled sufficiently many attempts with teammates that a distribution over common policies can be accurately estimated. In practice, such ideal circumstances may not occur; furthermore, it is of interest to find coordination techniques that are robust to inaccurate priors. Albrecht et. [8] compared baseline priors (uniform, random) with priors constructed under functions over individual agent payoffs. For example, a Utility prior weights teammate strategies proportionally to the payoff for the *coordinating agent*, while a Stackelberg prior assigns probabilities proportional to the payoff for the *teammate*. The authors observed that the best performing prior depended significantly on both the true teammate type and the depth of the coordinating agent's lookahead. Priors are an understudied element of ad hoc teamwork, as most existing work elects to use a uniform prior [3, 12, 10, 122, 120] while some prefer an experience-based prior[1] [14].

[1]The experiments in Chapter 11 construct a prior from past experience.

As a counterpart to the selection of initial beliefs, revising beliefs through observing teammate behavior, i.e. calculating the *posterior* belief distribution, has been a highlighted topic among much existing work [12, 14, 7, 122]. In typical inference problems, Bayes' rule ($\Pr(\mu_i|a_i) \propto \Pr(a_i|\mu_i)\Pr(\mu_i)$) is used to compute posterior distributions. However, consider an agent coordinating with a teammate acting under an entirely novel policy (μ^*) not covered by the agent's belief distribution, i.e. $\Pr(\mu^*) = 0$. It is possible that upon observing actions such that $\Pr(a_i|\mu_i) = 0 \ \ \forall \mu_i$, each modeled teammate policy is assigned a posterior probability of 0, causing a collapse of the belief distribution and leaving the agent unable to use observations to make predictions. Two common approaches circumvent this problem. First, choosing a prediction strategy such that $\Pr(a_i|\mu_i) > 0 \ \ \forall a_i$ ensures non-zero posteriors. Barrett et al. [12] manually overrides observations that drop out teammate models, electing to keep the previous probability instead. In Chapter 5, we use an exponentiated loss function, which is non-zero for mispredictions yet favors correct predictions in the posterior.

A second and more common approach is to adopt an alternative posterior formulation such that $\Pr(\mu_i|a_i) > 0$ for any observation. Barrett et al. [14] motivated the continued use of models of teammate behavior that only occasionally make mispredictions, necessitating a more lenient posterior update. For this purpose, the authors adopted a belief revision approach based on the polynomial weights algorithm (PWA) [23], which computes a posterior as a weighted mixture of the most current posterior, $\Pr_t(\mu_i)$, and the previous posterior, $\Pr_{t-1}(\mu_i)$. In a similar manner, we adopt a procedure in Chapter 5 based on the mixing of past posteriors [24], which can be viewed as a generalization of PWA in that any mixture of past posteriors, $\Pr_0(\mu_i), \ldots, \Pr_{t-1}(\mu_i)$, can be used in the posterior update. Moreover, Albrecht et al. [7] analyzed the conditions under which various posterior strate-

gies are guaranteed to converge to a teammate's correct type or type distribution, should the teammate changes strategies. Perhaps most noteworthy, the authors showed that under incorrect models, it is possible for a coordinating agent to mistakenly believe its policy will successfully coordinate with a teammate, potentially in an infinite cycle of incorrect decisions. The authors proposed adding the capacity to learn a teammate's true model in order to avoid this situation.

3.2 Learning

As one of the primary divergences from much of the traditional MAS literature, the assumption of heterogeneous team compositions manifests an appreciably distinct context for multiagent learning (MAL)[29]. The conclusions from the evaluation of MAL algorithms under typically homogeneous contexts may not hold in ad hoc settings, as crucial assumptions may be violated, particularly if teammates do not themselves learn or act according to an expected decision protocol. Despite this possibility, traditional MAL have been shown to be effective in ad hoc settings [5], though none of the methods tested were shown to have demonstrably superior performance than the others. Model-based techniques, in which predictive behavior models are learned for individual teammates, can learn accurate predictive models by adopting a classification perspective, using techniques such as decision trees [12, 14] and nearest-neighbor approaches [11]. Furthermore, transfer learning, the transfer of knowledge from a previously learned task to a new task, holds promise for ad hoc settings where current teammates may behave similarly to past collaborators for which an agent has previously learned models. Barrett et al. [13] adopted a weighting approach for reusing episodes of past experience, bootstrapping current observations of a new teammate with data from similar teammates when training a new behavior model.

3.3 Communication

Communication in ad hoc teams has received relatively light treatment. Barrett et al. [10] considered a communicative scenario wherein two agents participate in a form of the multi-armed bandit problem, a well-studied example in sequential decision making. It has been extended to ad hoc team settings to examine how information can be conferred to a teammate via demonstration as well as explicit communication. In the communicative, ad hoc version, an agent has the option to broadcast its last observation, the mean of a given arm, as well as a suggestion for which arm its teammate should pull. The communicative acts address the partially observable information within the domain—i.e. the payoff distributions of the bandit arms—but do not consider or aid in the process of resolving uncertainty in the teammate's future behavior, aside from making suggestions.

The Standard Platform League of the RoboCup Soccer division has recently featured a drop-in player competition [82, 47]. Teams of five members are drawn from the set of submissions. This domain is unique in that it features cooperative behavior with teammates and adversarial behavior against a separate team of ad hoc agents. Players were allowed basic communication using a standard set of messages regarding the positions of players and the ball, time since the ball was last seen, and the agent's intended role. However, not all teams utilized this ability; some teams reported that their robots disregarded most or all of the information received from teammates. In Chapter 6, we discuss existing work in multiagent communication for traditional multiagent systems, from which we motivate improved utilization of communication in ad hoc team settings.

3.4 Summary

This chapter has provided an overview of ad hoc teamwork literature pertaining to the core challenges faced by an ad hoc coordinating agent. As an overview of how the individual works differ from one another and also from this thesis, we have composed a comparison of techniques used in Table 3.1.

Reference	Domain	Prior	Posterior	Learning	Communication
Agmon [3]	Generated	Uniform	N/A	N/A	N/A
Barrett et al. [10]	Multi-armed Bandit	Uniform	Bayes	Indiv. Models	Bandit Info, Suggestions
Barrett et al. [12]	Multiagent Pursuit	Uniform	Bayes	Indiv. Models	N/A
Barrett et al. [14]	Multiagent Pursuit	Sampled	Polynomial Weights	Indiv. Models	N/A
Barrett et al. [11]	Half Field Offense	Sampled	Polynomial Weights	Indiv. Models	N/A
Albrecht et al. [5]	Repeated Games	N/A	N/A	Various (MAL)	N/A
Albrecht et al. [7]	N/A	N/A	Various	Type Distribution	N/A
Albrecht et al. [8]	Repeated Games	Various	Product	Type Distribution	N/A
Chapter 5	Multiagent Pursuit	Uniform	Mixture of Past Posteriors	Type Distribution	N/A
Chapters 7-11	Multiagent Pursuit	Sampled	Bayes	Type Distribution	Policy Info

Table 3.1: Related work summarized by techniques used for managing beliefs, learning, and communication.

Chapter 4

Multiagent Pursuit

Throughout this work, we will test various teamwork strategies using a variant of the multiagent pursuit problem, initially introduced in [19]. The multiagent pursuit problem—often referred to as a pursuit-evasion problem—has become a standard test domain for work in cooperative multiagent systems [148] and has been represented in many forms, depending on choices such as discrete or continuous state and action space, full or partial observability observability, as well as time constraints for making decisions [64, 145]. Regardless of these details, the premise of the problem is singular: a team of pursuers must track down and capture one or more evading agents.

Typically, applications to the multiagent pursuit problem have featured the tradition approach to MAS, wherein all agents share identical decision-making procedures. The problem, then, is one of sharing information and coordinating the actions of decentralized, autonomous team members. For the latter purpose, multiagent reinforcement learning has demonstrated the capacity for agents to learn coordination strategies [65] over many trials. Barrett et al. [12, 13, 14] extended this concept to ad hoc teamwork by demonstrating the ability for an agent to learn models of new teammate behavior using transfer learning. Rather

than learn one policy with a fixed team, an agent bootstraps learning with a new team by reusing models of behavior from past experience. Similarly, in a two pursuers/multiple evaders variant of the pursuit domain, Macindoe et al. [83] used an approach similar to that outlined in Chapter 2, wherein the strategy of a teammate is inferred from a set of modeled teammate *types*[1]. Both approaches to pursuit with unknown teammate strategies illustrated the need for better modeling of teammates, showing high success when the teammate follows a known strategy but requiring mechanisms for handling novel strategies (learned models in [12, 13, 14]; noisy predictions in [83]). For this reason, we will examine modeling of non-stationary behaviors (Chapter 5) as well as communication (Chapters 7-11) within this domain.

In contrast to the work of Barrett et al. [12, 13] where a team of four agents pursues a single evader, we will use the two-agent version from [83]. In this variant, an agent must collaborate with one teammate to capture one of the robbers, but there is uncertainty as to which is currently being pursued by the teammate. In order for the agents to successfully complete the game, they must simultaneously be present within the cell of an evader during a turn. If only one of the agents enters the evader's cell, the evader may slip by and flee. Furthermore, as the evader flees from the nearest pursuing agent, the pursuers must coordinate their approaches from distinct directions. Accurate predictions of the teammate's behavior, then, are key to successful coordination. Figure 4.1 shows five maze configurations used in [83], differing in layout, the number of robbers, and the inclusion of one-way doors, which can punish poor action selection by lengthening paths to targets as well as by trapping agents.

[1]The distinction between inference over teammate policies and teammate types is commonly that of quantity, where types implies a relatively small set [83, 13, 14, 116, 121] compared to the full policy space, or of stochasticity, where types can modeled as probabilistically choosing between policies [7, 122]

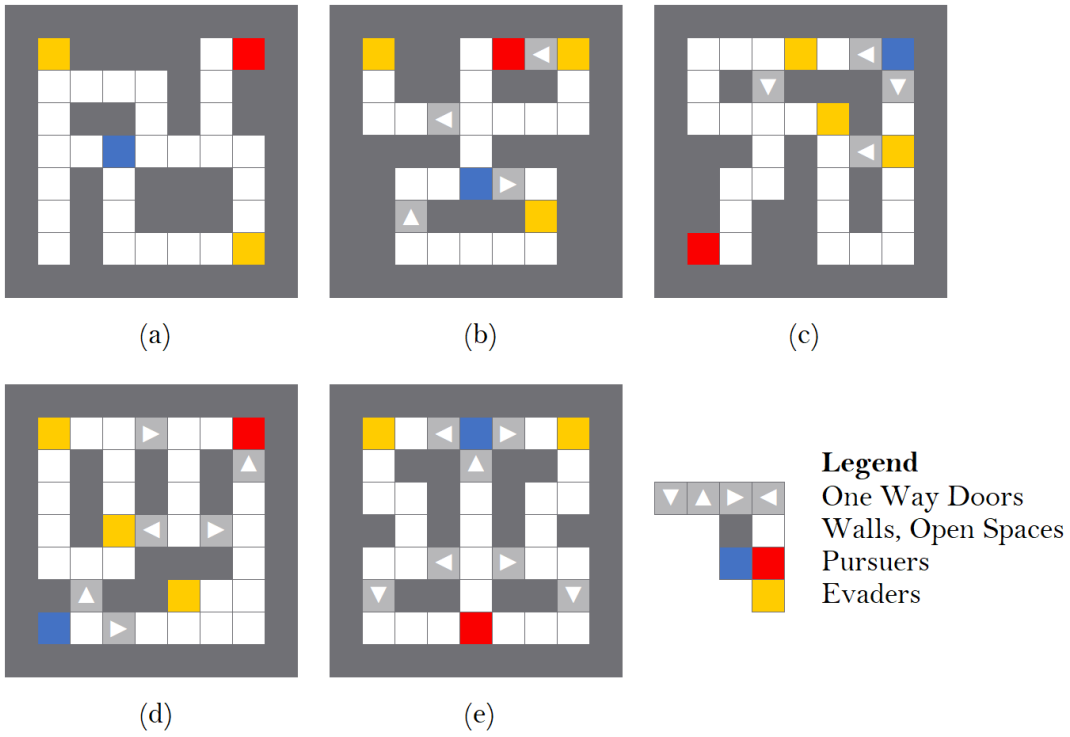

Figure 4.1: Mazes for the two-agent pursuit domain, as used in [83].

The domain proves challenging for multiagent planning due to the size of its state space. With two agents and two or three targets, this variant has a branching factor in the range of $2^4 = 16$ to $5^5 = 3,125$ each turn, depending on which cells the pursuers and evaders are located. While the behavior of the evaders is known to the pursuers, the individual policies of the pursuers are unknown. Given the partial observability resulting from uncertainty of the teammate's strategy, the domain is large enough that optimal POMDP planning is intractable.

4.1 Setup

The discrete, two-agent, turn-based pursuit problem operates as follows:

- All pursuers and evaders begin in locations as designated on a given maze,

for example, as shown Figure 4.1.

- States are given by the maze layout, positions of the pursuers and evaders, as well as the current round number.

- At each turn, the pursuers select individual actions from $\{\uparrow, \downarrow, \leftarrow, \rightarrow\}$. Available actions are constrained by the presence of walls and one-way doors, such that agents may not select an action that would result in moving into a wall or through a one-way door in an illegal direction.

- Evaders likewise select actions from $\{\uparrow, \downarrow, \leftarrow, \rightarrow\}$. The policies of the evaders select actions maximizing the minimum distance to a pursuer, i.e. $\pi_e(s) = \arg\max_{a_e \in A_e} \left[\min_p D(e, p, a_e)\right]$, where D is the distance between e and p after taking action a_e.

- The task ends when either a round limit is reached (specified by maze/application) or when both pursuers have moved into a cell of one or more evaders, i.e. $\exists e \in E \ : \ loc(e) = loc(p_1) = loc(p_2)$.

- When a terminal state is reached, the team receives a reward determined by the particular application. Typically this is a flat reward ($R = 100$) or linear in the number of turns ($R = 100 - t$).

Chapter 5

Case Study: Non-stationary Strategy Approximation for Unknown Teammates

Effective teamwork relies on the coordination of individual team members which, in turn, requires the team to have formed a consensus on the task at hand. In many settings, a consensus could be reached through communication, where teammates could weigh in on goals, plans, and other relevant information. Communication, however, may not always be guaranteed, particularly if two agents were designed separately and without any shared information regarding one another. Under these conditions, efficient teamwork depends not only on an agent's ability to infer but also its adaptability to change. Traditionally, ad hoc research has assumed that the behavior of an unknown teammate is stationary, meaning that a teammate will stick to a single individual policy. In this chapter, we consider a broader case where an unknown teammate may change its behavior

The results of this chapter are presented in [121, 122].

unexpectedly.

We introduce an approach, Responsive Action Planning with Intention Detection (RAPID), for updating beliefs over an agent's policy with bounds on the number of observations necessary to identify a change in teammate behavior. In many domains, it may be inaccurate to assume a teammate will stick to a single goal or strategy throughout a collaborative task, especially when provided with an incentive to switch, whether it be an easier route to a goal or simply a more appealing one. An ideal team agent should not only be able to assist its teammate in achieving its goals but also be flexible in its capacity to account for changes in teammate behavior. In ad hoc team settings, we must consider the potential for a teammate to adapt an entirely novel policy from the observer's perspective. To account for this complexity yet retain a desirable degree of practical application, we propose an adjustment to the belief revision process such that potential alternate goals are kept at relevant minimum likelihoods. As a result, RAPID bounds the number of observations required to identify a switch in a teammate's pursued goal at the expense of being sensitive to inconsistent or noisy behavior.

As in related work in ad hoc teamwork, [13, 14, 10], RAPID models the planning space as a partially observable Markov decision process (POMDP). However, in contrast to the decision problem outlined in Chapter 2, this chapter proposes an approach for scenarios in which a teammate's behavior appears non-stationary from the perspective of a coordinating ad hoc teammate. Over the course of many observations, the agent's beliefs may favor different models in different periods of time. This provides a simplified representation of how an agent may adopt a new plan on the fly according to an unmodeled, unknown, underlying preference. However, when enough observed evidence indicates a switch change in behavior, it is vital that a coordinating agent adapt its beliefs responsively.

5.1 An Alternative Approach to Belief Revision

An important aspect of planning in a partially observable scenario is the ability to refine a set of beliefs corresponding to the hidden information in the state space. This is completed through inference after observing some aspect of the world or action of an agent. As by definition, a POMDP is in part defined by a set of probabilities for observations made in each potential state. Traditionally, beliefs are revised using the observation history and Bayes' theorem:

$$\Pr(\mu_i|a_i) = \frac{\Pr(a_i|\mu_i)\Pr(\mu_i)}{\Pr(a_i)} \tag{5.1}$$

This approach, however, requires assumptions that do not hold in ad hoc teamwork. Primarily, it is assumed that the true model of a teammate is contained within the set covered by the belief distribution, b. For this to be true, it may require maintaining a belief distribution over an intractably large policy space. However, maintaining beliefs on a relatively small set of teammate policies runs the risk of ruling out each of the known models ($\Pr(a_i|\mu_i) = 0,\ \forall i$), invalidating the beliefs. Therefore, we are motivated by techniques which model the prediction problem under similar constraints, wherein a weighted set of *experts* are used to make predictions. As the concept of an agent with shifting priorities has natural similarities to shifting experts/online-learning problems, we borrow the concept of modifying the belief revision step by adding a mix of past posteriors, as described in [24].

As a first step in this new belief revision construction, we make use of a simple

binary loss function, L for model or policy μ_i given an observed action, a_i:

$$L(\mu_i) = \begin{cases} 1 & \mu_i(s) = a_i \\ 0 & \text{otherwise} \end{cases}$$

We then estimate the action likelihoods using

$$\Pr(a_i|\mu_i) \propto e^{-L(\mu_i)},$$

yielding a belief revision formulation of

$$\Pr(\mu_i|a_i) \propto e^{-L(\mu_i)} \Pr(\mu_i). \tag{5.2}$$

Finally, we modify the revision step by adding in a weighted portion of the agent's *initial* belief distribution, as explored by [24].

$$\Pr'_t(\mu_i|a_i) = \beta \Pr_0(\mu_i) + (1-\beta) \Pr_t(\mu_i|a_i) \tag{5.3}$$

This modified approach bears much resemblance to the polynomial weights algorithm[1] [23] as used previously in [13]. In the latter approach, using a polynomial weight slows the belief convergence to a particular teammate model such that no model is discarded prematurely; however, given a sufficiently long series of observations supporting one model, the probabilities can still diverge substantially, such that switching to favor an alternate model can require a correspondingly large number of observations. In contrast, by mixing the updated belief vector with the initial belief vector, we are able to enforce upper and lower bounds on

[1]The polynomial weights algorithm is a specific form of the mixing of past posteriors method. It takes the form $\Pr'_t(\mu_i|a_i) = \beta \Pr_{t-1}(\mu_i) + (1-\beta) \Pr_t(\mu_i|a_i)$. Our method incorporates the initial prior rather than the most recent posterior.

the possible values of the agent's belief probabilities.

The mixing parameter is constrained by $0 \leq \beta \leq 1$. When $\beta = 0$, the updated probability takes the form of Bayes' theorem, while a value of $\beta = 1$ results in a stationary probability equal to the initial probability assigned. In this context, β is a domain-dependent hyperparameter, the selection of which may take into account factors such as the relative consistency of teammate behavior or a desired number of observations required to realign an agent's beliefs.

5.2 Parameter Tuning

Choosing an appropriate value of β is crucial for responsively identifying hidden state changes. Under a traditional belief revision approach, identification of hidden state transitions can require long sequences of observations. In fact, under certain conditions, identification of a goal switch can be linearly dependent on the number of observations supporting a previous model of behavior. This is particularly undesirable for domains with significantly many observations before a switch occurs.

Theorem 1. *Let* $\Pr_{t_m+t_n}(\mu_i)$ *and* $\Pr_{t_m+t_n}(\mu_j)$ *be the probabilities of two teammate policies after* t_m *and* t_n *observations supporting* μ_i *and* μ_j *respectively, using the update procedure of Equation 5.2. Then,*

$$\Pr_{t_m+t_n}(\mu_j) \geq \Pr_{t_m+t_n}(\mu_i) \implies t_n \geq \ln \frac{\Pr_0(\mu_i)}{\Pr_0(\mu_j)} + t_m$$

Proof. The property is shown by performing the update for both states, resulting

in

$$\Pr_{t_m+t_n}(\mu_i) = \eta \Pr_0(\mu_i) e^{-t_n}$$
$$\Pr_{t_m+t_n}(\mu_j) = \eta \Pr_0(\mu_j) e^{-t_m}$$

where η is a product of normalizing factors. The result follows directly.

The alteration proposed in Equation 5.3 does not have a closed form solution for the number of steps required for a state switch to be identified. However, empirically we observe that tuning of β creates an upper bound for the number of steps required under ideal observations. Figure 5.1 depicts the effect of various tested β values under thirty potential goals and observations which only support a unique goal. Despite an increase in the steps a teammate pursues the first goal, the required number of observations supporting a second goal to converge to the appropriate belief is bounded.

Given that the modified belief revision approach can bound the number of required observations to any arbitrary number, selecting a value for β has the trade-off between responsiveness—quickly reweighting to a more correct model—and susceptibility to noise, mistakenly favoring a model after only a few observations. A series of inaccurate observations or observations of actions by an agent imperfect in its pursuit of a goal can lead belief convergence to the wrong target goal. We will refer to these observations as *noisy* observations. In these circumstances, tuning β is dependent on the likelihood of such noise in the domain tested.

We define noisy observations as those supporting any subset of goals not including the true underlying goal currently being pursued by an agent. In the worst case, observations support exactly one incorrect goal, i.e. $\exists! \mu_i \quad L(\mu_i) =$

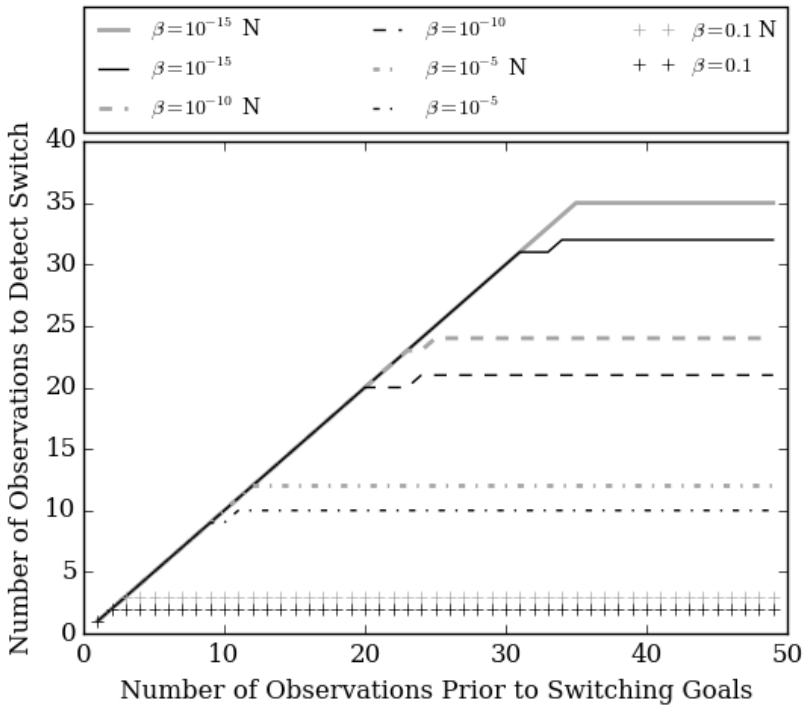

Figure 5.1: Number of observations required for the belief distribution to align with the corresponding goal as a function of the number of observations supporting the prior goal. Lines denoted with N correspond to the normalized belief revision approach. All remaining cases are not normalized.

0, and $\mu_{\text{i}} \neq \mu^*$ where μ^* is the true policy. If multiple successive noisy observations occur supporting a single incorrect goal, the belief distribution can converge to the incorrect state.

Consider the case where a domain has a noise rate r. The probability of a number of successive noisy observations, K, forms a geometric distribution with $\Pr(K = k) = r^k(1 - r)$. The expected length of a sequence of noisy observations, then, is given by $\mathbb{E}[K] = \frac{1}{1-r}$. It is reasonable to constrain β such that the number of required observations to identify a switch in underlying state to $n \geq \left\lceil \frac{1}{1-r} \right\rceil$.

Due to the normalization of probabilities Equation 5.3, choice of β depends on the noise rate tolerance as well as the number of alternative goals, as normalization

considers the probabilities of all possible goals involved. Such normalization is necessary for decision-theoretic reasoning, as the sum of probabilities of underlying states must sum to one. The non-normalized version, given by Equation 5.4, is useful for approximation of the relative quantities of the normalized case. Figure 5.1 depicts a comparison of the two versions, with N denoting the normalized cases.

$$\Pr_t^\star(\mu) = \beta \Pr_0(\mu) + (1-\beta) \Pr_{t-1}^\star(\mu) e^{-L_{t-1}(\mu)} \tag{5.4}$$

$$= \beta \Pr_0(\mu) + \beta \Pr_0(\mu) \sum_{i=1}^{t-1} (1-\beta)^i e^{-\sum_{j=t-i}^{t-1} L_j(\mu)}$$

$$+ (1-\beta)^t \Pr_0(\mu) e^{-\sum_{k=0}^{t-1} L_k(\mu)} \tag{5.5}$$

The expanded case, represented by Equation 5.5, allows for the direct calculation of the upper and lower bounds of these pseudo-probabilities. In the limit, as $t \to \infty$ and $L_t = 1 \ \forall t$, the result settles at $\Pr_\infty^\star(\mu) \geq \frac{\beta \Pr_0(\mu)}{1-e^{-1}(1-\beta)}$. Similarly, expanding with loss $L_t = 0 \ \forall t$ yields the upper bound $\Pr_\infty^\star(\mu) \leq \Pr_0(\mu)$. With these bounds established, we can compute the number of steps required for a state with minimal probability to succeed one with maximal probability when comparing the likelihoods of the two policies.

Lemma 1. *Let μ_a, μ_b be drawn from an initially uniform distribution, such that $\Pr_0(\mu_a) = \Pr_0(\mu_b)$. At time t, let $\Pr_t^\star(\mu_a) < \Pr_t^\star(\mu_b)$. In order to guarantee $\Pr_{t+n}^\star(\mu_a) > \Pr_{t+n}^\star(\mu_b)$, it must be that $\beta > 1 - \left(\frac{1}{1+e^{-n}}\right)^{\frac{1}{n}}$.*

Proof. By the bounds established earlier, observe that $\Pr_t^\star(\mu_a) \geq \frac{\beta \Pr_0(\mu_a)}{1-e^{-1}(1-\beta)}$ and $\Pr_t^\star(\mu_b) \leq \Pr_0(\mu_b)$. The proof follows by expanding Equation 5.5 n steps beyond

t for each policy.

$$\begin{aligned}
\Pr^{\star}_{t+n}(\mu_a) =& \beta \Pr_0(\mu_a) + \beta \Pr_0(\mu_a) \sum_{i=0}^{n-1} (1-\beta)^i e^{-\sum_{j=t+n-i}^{t+n-1} L_j(\mu_a)} \\
&+ (1-\beta)^n \Pr^{\star}_t(\mu_a) e^{-\sum_{k=t}^{t+n-1} L_k(\mu_a)} \\
\geq& \beta \Pr_0(\mu_a) + \beta \Pr_0(\mu_a) \sum_{i=0}^{n-1} (1-\beta)^i e^{-\sum_{j=t+n-i}^{t+n-1} L_j(\mu_a)} \\
&+ (1-\beta)^n \frac{\beta \Pr_0(\mu_a)}{1 - e^{-1}(1-\beta)} e^{-\sum_{k=t}^{t+n-1} L_k(\mu_a)}
\end{aligned}$$

$$\begin{aligned}
\Pr^{\star}_{t+n}(\mu_a) =& \beta \Pr_0(\mu_b) + \beta \Pr_0(\mu_b) \sum_{i=0}^{n-1} (1-\beta)^i e^{-\sum_{j=t+n-i}^{t+n-1} L_j(\mu_b)} \\
&+ (1-\beta)^n \Pr^{\star}_t(\mu_b) e^{-\sum_{k=t}^{t+n-1} L_k(\mu_b)} \\
\leq& \beta \Pr_0(\mu_b) + \beta \Pr_0(\mu_b) \sum_{i=0}^{n-1} (1-\beta)^i e^{-\sum_{j=t+n-i}^{t+n-1} L_j(\mu_b)} \\
&+ (1-\beta)^n \Pr_0(\mu_b) e^{-\sum_{k=t}^{t+n-1} L_k(\mu_b)}
\end{aligned}$$

The observations supporting μ_i result in losses of $L(\mu_i) = 0$ and $L(\mu_j) = 1$.

$$\begin{aligned}
\Pr^{\star}_{t+n}(\mu_a) \geq& \beta \Pr_0(\mu_a) + \beta \Pr_0(\mu_a) \sum_{i=0}^{n-1} (1-\beta)^i + (1-\beta)^n \frac{\beta \Pr_0(\mu_a)}{1 - e^{-1}(1-\beta)} \\
\Pr^{\star}_{t+n}(\mu_a) \leq& \beta \Pr_0(\mu_b) + \beta \Pr_0(\mu_b) \sum_{i=0}^{n-1} (1-\beta)^i e^{-i} + (1-\beta)^n \Pr_0(\mu_b) e^{-n}
\end{aligned}$$

As we are interested in $\Pr^{\star}_{t+n}(\mu_a) > \Pr^{\star}_{t+n}(\mu_b)$, we can compare the lower bound for the former with the upper bound for the latter.

$$\begin{aligned}
&\beta \Pr_0(\mu_a) + \beta \Pr_0(\mu_a) \sum_{i=0}^{n-1} (1-\beta)^i + (1-\beta)^n \frac{\beta \Pr_0(\mu_a)}{1 - e^{-1}(1-\beta)} \\
&\geq \beta \Pr_0(\mu_b) + \beta \Pr_0(\mu_b) \sum_{i=0}^{n-1} (1-\beta)^i e^{-i} + (1-\beta)^n \Pr_0(\mu_b) e^{-n}
\end{aligned}$$

Given that $\Pr_0(\mu_a) = \Pr_0(\mu_b)$, we can divide out the common term. Furthermore, as $(1-\beta) < 1$ and $e^{-1} < 1$, the summations are finite geometric sums and can be evaluated directly.

$$\beta\frac{1-(1-\beta)^n}{1-(1-\beta)} + (1-\beta)^n\frac{\beta}{1-e^{-1}(1-\beta)} \geq \beta\frac{1-(1-\beta)^n e^{-n}}{1-(1-\beta)e^{-1}} + (1-\beta)^n e^{-n}$$

Algebraic manipulation achieves the final result.

With a lower bound for β established, it remains only to choose an acceptable n, which we have discussed previously. Experimental evaluation of two β values for a set noise rate is described in the next section.

5.3 Evaluation

We evaluate the proposed belief revision modification in the two agent pursuit domain outlined in the previous chapter. Two notable works exist in this and a similar domain. Macindoe et al. [83] introduced Cops and Robbers as a domain for testing sequential planning for assistive agents under partial-observability with respect to a teammate's goal; however, the teammate agent in the evaluation chose a single target at the start and never switched for the duration of the game. Nguyen et al. [96] used a similar game, Collaborative Ghostbuster, and modeled the switching of targets with a simple, fixed probability. Conversely, rather than compute a coordinating policy under the POMDP representation, the authors propose dividing the task into individual worlds, one for each potential evader. The agent maintains a belief and policy for each world, selecting its next action as that maximizing the expectation over possible worlds. In contrast to the POMDP approach, this technique does not account for future observations and corresponding belief states; it merely maintains its current beliefs. A consequence

of this strategy is that it can favor remaining near two less likely targets rather than pursue the most likely target, as pointed out in [83]. In contrast to these two works, our approach plans in belief space and does not explicitly model the unobserved goal transitions, yet it can identify changes in behavior quickly if and when they occur.

5.3.1 Agents

For our tests, we implemented three teammates whose goal remains uncertain to the coordinating agent. The teammates behave as follows:

- A* - Pursues the closest evader at the start of the game and never switches targets.
- Switch Once - Switches targets at a fixed point in the game, on the eighth turn.
- Probabilistic - Switches targets to evader, $e_j \in E$, with probability proportional to its distance, D:

$$\Pr(e_i \rightarrow e_j) = \frac{0.2|E|\ D(e_j)}{\sum\limits_{e \in E} D(e)} \tag{5.6}$$

All teammates move toward the selected target using A* path planning, with 10% noise in their actions. A noisy action is randomly selected from the set of possible actions that would not pursue the active target.

For the ad hoc agent, we implement several Monte Carlo Tree Search (MCTS) [34] agents using Upper Confidence Bounds applied to Trees (UCT) [75] as the action selection strategy, varied by how they model the unknown teammate:

- UCT - Performs multiagent planning with UCT. This agent assumes its teammate will plan and behave in an identical manner to itself and arrive at an identical joint policy.

- Bayes - Plans using UCT for its own actions but uses a belief distribution over possible teammate goals. Updates beliefs according to Equation 5.2.

- RAPID - As above, but updates its belief distribution according to Equation 5.3.

- Limited Oracle - Plans with perfect knowledge of the teammate's *current* target. However, it cannot predict future goal changes.

5.3.2 Tests

Each pair of teammate and coordinating agent participate in one hundred trials of each maze featured in Figure 5.2, repeated here for convenience. Steps taken to complete the game, beliefs of applicable agents, and targets of the teammates are logged for analysis. We allow each UCT-based agent one hundred game simulations per turn, with root parallelization [35] across four cores.

To emphasize the effect of tuning, we use two versions of our RAPID agent, each with a different value for β. Given the 10% noise rate in our experiment and the geometric distribution of expected successive noisy actions, we observe that 99.9% of groups of successive noisy actions are of length 2 or fewer. Choosing $n = 3$, then, gives us a lower bound for choice of $\beta = 0.016$ from Lemma 1. For a less conservative tuning, the remaining RAPID agent uses $\beta = 0.85$ for enhanced responsiveness at the risk of susceptibility noise.

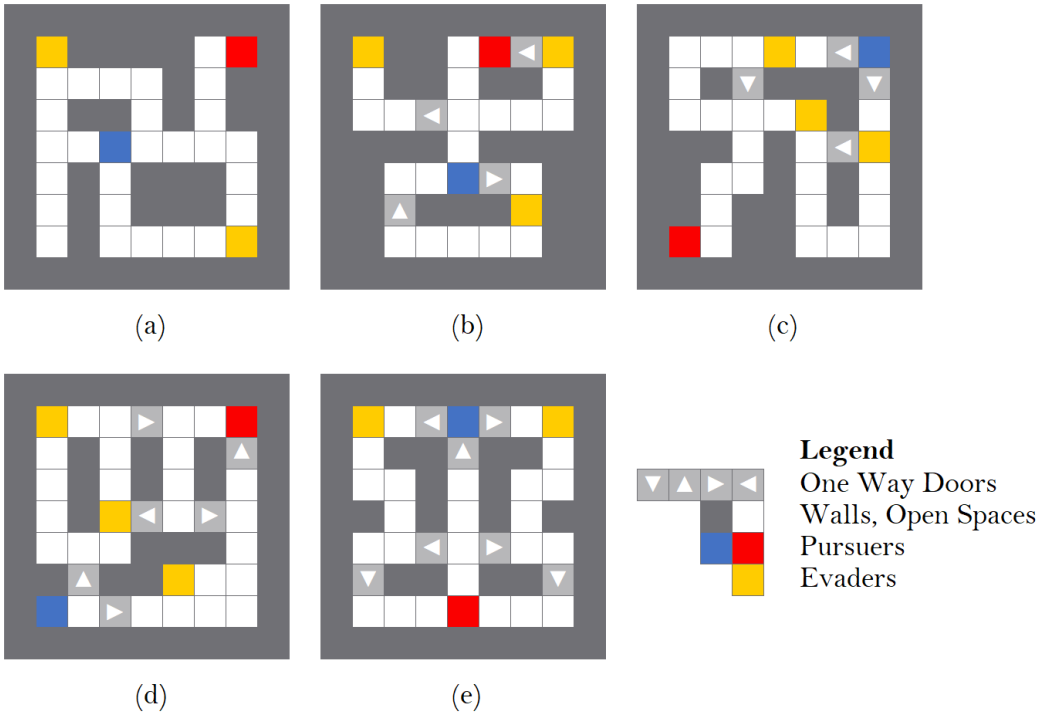

Figure 5.2: Mazes for the two-agent pursuit domain, as used in [83]. The coordinating agent is represented in blue, with the teammate in red.

5.4 Results

We compare the performance of the RAPID agent against the agent which revises its beliefs with a traditional Bayes update. The plain UCT agent provides the base level of performance we would expect with any of the UCT-based agents, while the limited oracle agent demonstrates the potential for further improvement. It should be noted that the results of the limited oracle agent could be unattainable, as the agent has access to the teammate's true target. Furthermore, as the agent is still primarily a sampling-based agent and also lacks the foresight of future target transitions, it should not be mistaken for an optimal ad hoc agent.

Three metrics serve to evaluate our proposed approach: the average number of observations required to revise an agent's beliefs to the appropriate target, the

percentage of steps in our tests in which an agent has correctly identified the target, and the average number of steps required to capture a target in the maze.

5.4.1 Belief Recovery

Table 5.1 reports the number of times in 100 trials the teammate switched targets as well as the average steps required for the agents to identify the change. The base UCT and limited oracle agents are omitted as they do not possess a belief system. The A* teammate is similarly absent, as it never switched targets.

With respect to belief recovery time, the RAPID agent with the conservative tuning of β only outperforms the Bayes agent in one of ten test cases ($\alpha = 0.01$). Between noisy actions and those supporting potentially two or more targets, the RAPID ($\beta = 0.016$) agent could not utilize the bounded convergence time to significant effect.

The second RAPID agent, however, outperforms the Bayes agent in six of the ten relevant test cases, with comparable performance in the remaining four cases. In the instances of improvement, the agent was able to detect a switch with fewer observations on average than the Bayes agent, resulting in an average gain of nearly seven turns in one test case.

5.4.2 Accuracy

With a shorter time to converge to a pursued goal, it is natural to expect an increase in accuracy of the predicted goal. For this metric, steps where the correct target probability is equal to that of another target are considered ambiguous and are counted as an incorrect identification. This explains a portion of the observed low accuracies, particularly as the first few steps in each game are not sufficient to distinguish targets.

		Bayes		$\beta = 0.016$			$\beta = 0.85$		
	Teammate	n	Average	n	Average	p	n	Average	p
a	SwitchOnce	100	5.04	96	**3.76**	<0.001	100	**1.00**	<0.001
	Probabilistic	269	4.42	400	5.14	0.096	363	**2.78**	<0.001
b	SwitchOnce	99	18.04	92	18.40	0.457	92	23.30	0.079
	Probabilistic	369	12.96	326	17.83	<0.001	472	11.72	0.145
c	SwitchOnce	94	7.57	66	6.53	0.248	67	9.06	0.221
	Probabilistic	454	12.92	502	14.25	0.128	356	**8.10**	<0.001
d	SwitchOnce	100	15.87	100	14.55	0.347	100	11.75	0.085
	Probabilistic	557	15.48	644	14.19	0.133	532	**9.45**	<0.001
e	SwitchOnce	100	18.7	61	18.18	0.443	100	**11.79**	0.007
	Probabilistic	506	8.85	356	7.86	0.132	396	**6.09**	<0.001

Table 5.1: Average actions observed before teammate's true target is most likely in agent's belief distribution. Bold values indicate significant results over the Bayes agent ($\alpha = 0.01$).

		Bayes		$\beta = 0.016$			$\beta = 0.85$		
	Teammate	n	% Correct	n	% Correct	p	n	% Correct	p
a	A*	2188	17.69	2226	17.83	0.448	1617	**23.69**	<0.001
	SwitchOnce	3201	71.88	3333	**78.97**	<0.001	2794	**80.96**	<0.001
	Probabilistic	2159	57.94	2634	**62.34**	<0.001	2476	**64.01**	<0.001
b	A*	3792	26.85	3796	**48.97**	<0.001	4181	25.52	0.089
	SwitchOnce	3771	39.30	4574	**46.55**	<0.001	5100	34.27	<0.001
	Probabilistic	3712	33.14	4280	30.72	0.010	4029	**38.07**	<0.001
c	A*	2671	40.43	2406	22.94	<0.001	2522	33.51	<0.001
	SwitchOnce	3472	60.77	2521	52.52	<0.001	2689	51.95	<0.001
	Probabilistic	3533	42.37	3960	**46.31**	<0.001	2763	**47.09**	<0.001
d	A*	2516	14.63	2708	**31.50**	<0.001	2962	**26.00**	<0.001
	SwitchOnce	6358	49.53	5223	**57.15**	<0.001	4927	48.81	0.227
	Probabilistic	4412	45.42	5157	**50.69**	<0.001	5048	**57.81**	<0.001
e	A*	2527	14.88	1356	15.71	0.245	1939	13.67	0.125
	SwitchOnce	4480	35.71	2420	32.73	0.006	3562	**38.63**	0.004
	Probabilistic	3454	40.56	2536	42.59	0.058	2885	35.94	<0.001

Table 5.2: Percentage of steps with correct target identified by belief distribution. Bold values indicate significant results over the Bayes agent ($\alpha = 0.01$).

	Teammate	Bayes $\overline{steps}$	$\beta = 0.016$ $\overline{steps}$	p	$\beta = 0.85$ $\overline{steps}$	p	UCT $\overline{steps}$	Ltd Oracle $\overline{steps}$
	A*	21.88	22.26	0.323	**16.17**	<0.001	51.95	31.97
a	SwitchOnce	32.01	33.33	0.480	**27.94**	0.005	71.07	18.74
	Probabilistic	21.59	26.34	0.063	24.76	0.006	64.89	25.91
	A*	37.92	37.96	0.125	41.81	0.460	57.76	54.89
b	SwitchOnce	37.71	45.74	0.164	51.00	0.010	61.81	49.83
	Probabilistic	37.12	42.8	0.171	40.29	0.378	56.6	41.48
	A*	26.71	**24.06**	<0.001	**25.22**	0.004	58.11	28.74
c	SwitchOnce	34.72	**25.21**	<0.001	**26.89**	<0.001	67.58	35.88
	Probabilistic	35.33	39.60	0.309	**27.63**	0.002	71.78	31.74
	A*	25.16	27.08	0.174	29.62	0.068	69.65	39.94
d	SwitchOnce	63.58	52.23	0.010	**49.27**	0.001	84.08	75.97
	Probabilistic	44.12	51.57	0.058	50.48	0.043	81.05	42.39
	A*	25.27	**13.56**	<0.001	**19.39**	0.004	62.08	11.38
e	SwitchOnce	44.80	**24.20**	<0.001	**35.62**	<0.001	81.24	20.17
	Probabilistic	34.54	**25.36**	<0.001	**28.85**	0.009	73.21	19.7

Table 5.3: Average steps taken by the team to capture a target. Bold values indicate significant differences over the Bayes agent ($\alpha = 0.01$).

With regard to overall accuracy, the RAPID agents were found to be correct more frequently in the majority of scenarios. Both β levels had significant accuracy improvements over Bayes in eight test cases each. The Bayes agent outperforms the $\beta = 0.016$ agent in two instances and the $\beta = 0.85$ agent in four instances, as seen in Table 5.2. This loss of accuracy in the higher β value, particularly in cases shown to have significantly shorter belief recovery periods, demonstrates the susceptibility to noise, as was expected in the tuning of β.

5.4.3 Steps Taken

Table 5.3 shows the average number of turns required to complete each test case. The less responsive of the RAPID agents had significant improvements over the Bayes agent in five of the fifteen test cases. Furthermore, it no test cases did it perform significantly worse.

A higher β value, having reduced belief correction time and improved accuracy,

resulted in coordination time improvements in nine test cases. The Bayes agent only achieved a higher average score than the $\beta = 0.85$ agent in one case.

Additionally, results for the remaining tested agents are included for comparison. The base UCT agent, which assumes identical planning on the part of its teammate, demonstrates the benefit of accurate modeling, as it performed worse in every scenario than any other tested agent. The limited oracle version illustrates in a few test cases where there is still room for improvement. This is notably true in Maze e, where the oracle agent outperforms both RAPID agents. It should also be noted that in Maze b, knowing the correct target yet having no knowledge that the target may change led to poor performance, as the agent was susceptible to being trapped in the left corridor. A more comprehensive model of a teammate would take into account the likelihood of a change and the potential risks of taking actions toward the pursuit of the current target should a switch occur.

5.5 Discussion

Most existing work in ad hoc teamwork assumes a stationary behavior or goal for teammates. This chapter introduced a variation of the pursuit domain in order to evaluate approaches to working with a teammate whose goals and corresponding behavior can change periodically. Planning under our proposed changes to belief revisions allows an agent to quickly recognize and adapt to altered behavior indicative of a goal switch. Faster belief convergence to the correct goal boosts overall accuracy of the agent's predictions, which are directly leveraged in planning for better multiagent coordination. Tuning of the presented approach is likely specific to the domain but has been proven effective the in the various combinations of potential teammates and particular maps tested in this chapter.

Secondly, this initial empirical evidence suggests that reasoning quickly over a set of independent models may provide an acceptable approximation to modeling higher level reasoning of an unknown teammate, as long as its base behaviors are represented in the set of models. An optimal decision-theoretic agent would be capable of reasoning over teammate behavior as well as the likelihood of changes. However, under such possibilities, the considered state space grows drastically. We observe that a set of possible behaviors and a responsive belief revision approach can approximate a full model well.

Human-agent teamwork is one potential application of using a set of basic behavior models to approximate a high level decision process. Existing work for modeling human cognition often utilizes *theory of mind* concepts [150] or relies on learned or hand-authored models. If an agent is to assist a human in an environment that has clear potential goals and corresponding behaviors, our approach may prove advantageous. It is likely easier to design predictive models for simple goals, compared to more complex cognitive models. Furthermore, more responsive switching of tasks may be an acceptable response to the high-level decision making of the human teammate. It forgoes much computation on the larger body of tasks to be completed in favor of coordinating reactively. Naturally, this puts the agent in a supporting role while a human takes the lead in prioritizing goals.

Chapter 6

Motivation for Communicating Plans

As a brief recap, we are primarily interested in the performance of an ad hoc agent coordinating with one or more unknown teammates given an amount of prior experience[1] coordinating with other teams on a given task. At its core, the ad hoc teamwork problem is one of information deficiency, as policies of teammates can be varied among a population of potential teammates, and such policies are only indirectly observable during the act of coordination. This deficiency is addressed primarily through two means: leveraging past experience to inform a prior over teammate policies at the beginning of each trial and utilizing observed actions to narrow the space of likely policies through inference. Often, the diversity of teammate policies is directly related to the quantity of observations required to correctly and exactly infer a teammate's policy. However, ad hoc teamwork is not posed as a problem of exact classification but rather one of reward or utility maximization. It is often sufficient to narrow the space of policies to those which

[1] Or, alternatively, a prior distribution formed over teammate policies by other means, such as manual specification.

agree on a subset of the state space, for example those along the trajectory between the current state and some future goal state. If a set of potential policies align over these states, they are considered *behaviorally equivalent* [106]. In contrast, it may be that a teammate can choose between multiple policies that each achieve the team's goal regardless of the coordinating agent's individual policy, leaving the agent indifferent to the uncertainty of the choice. In this way, inference is a mechanism by which an ad hoc agent may improve its ability to coordinate, yet the necessity and effectiveness of inference on the success of coordination is ultimately a domain-dependent factor.

In Chapter 1, we proposed explicit communication of policy information as a potential solution for coordination under uncertainty of teammate policies. Here, we discuss the conditions under which non-communicative approaches struggle to achieve success.

6.1 Difficulties for Observation-based Inference

6.1.1 Imperfect Beliefs

Consider, as an example, the problem of possessing a prior over teammate policies that does not include the current observed policy, i.e. encountering a teammate whose strategy is novel to the agent. At the heart of this scenario is the problem of prior experience, as the agent has either never coordinated with such an agent or has never considered such a strategy probable. One such example of this is the non-stationary policy problem of Chapter 5, wherein a teammate that switches at some time from one policy to another, both of which are known but neither of which individually predicts the complete observed policy. Modifying the posterior belief update allows an agent to identify when a change in policy occurs;

however, such an approach does not consider the prediction of policy changes but instead relies on adapting to the new policy once a switch is detected. The implicit assumption of such approaches is that partially accurate models may retain useful predictive power despite occasionally making incorrect predictions. As discussed in [7], the ability for an incorrect belief distribution to succeed in this manner is dependent on the relation between the set of approximate models and the true teammate type space. The authors prove that under certain conditions in stochastic Bayesian Games, such approximations can still guarantee successful coordination. Nonetheless, there also exist conditions under which the modeling agent believes it is selecting a policy which coordinates correctly, while in truth, the agent's adapted policy repeatedly fails. The authors propose incorporating a means of learning a correct teammate policy in order to eventually counteract the conditions causing coordination failures.

6.1.2 Learning

The process of learning a teammate behavior model has been used across various work within the ad hoc teamwork community. Barrett et al. [12] employed decision trees in order to learn representations of teammate policies during the act of coordination. The authors reported successful predictive models learned using relatively few observations within the multiagent pursuit domain. This likely is in part due to the relatively small action space (per agent) as well as the consistent trajectories taken by agents, allowing for many similar states to map to the same action. The notion of grouping states according to their similarity is further discussed in [6], where *conceptual types* are introduced as methods of abstracting policy decisions over similar states in an effort to generalize observed behavior to predict actions in unobserved states. In domains in which states do not fall

into clear groupings or where policies are not consistent within such groupings, an agent may need to rely on significantly more observations—perhaps over repeated episodes of coordination—in order to learn a correct model. Such a requirement, however, may prove difficult in instances of teamwork with time constraints or limited observability.

One approach to supplementing small numbers of observations is to reuse data from past instances of coordination, as proposed in [14]. Adapting a concept from transfer learning, the authors proposed a method of learning a classifier by aggregating data from past experience with other teammates along with observations of the current teammates. As certain episodes of past experience may be more similar to the current coordinated effort than others, the proposed algorithm attempted a form of greedy weighting, tuning the weights of episodes to minimize the error of the classifier over observations of the current teammate. While the empirical tests demonstrated the effectiveness of the transfer learning approach, the quantity of observations of the current teammate played a large effect on the overall performance of the learned model. Only for observation set sizes of $1,000$ or $10,000$ were the results near optimal, with sets on the order of 10 or 100 performing significantly worse.

6.1.3 The Informativeness of Observations

In the approaches outlined thus far, we have discussed effectiveness in relation to the *quantity* of observations; however, in truth, the ability to successfully deduce a teammate's policy is dependent on the informativeness of observations available to the agent. It is necessary to observe divergences in policies among models of behavior in order to rule out models which will incorrectly predict future actions of the team. As discussed in [7], models with high overlap can cause

convergence to an incorrect type distribution under certain posterior update functions. Furthermore, consider a scenario where models each predict an identical series of teammate actions until some time t, upon which they diverge at a critical juncture where the ad hoc agent must select an important coordinating action. As the previous $t-1$ observations have not provided sufficient information to infer the appropriate course of action, the coordinating agent is left with uncertainty despite the potentially arbitrarily large number of observations it has witnessed. A similar situation develops in domains where agents may have limited or no observability with regard to their teammates actions, such as when working in separate locales, as in rescue robotics or unmanned aerial vehicle domains. Perhaps surprisingly, little attention has been given to the quality of observations in ad hoc teamwork, as the informativeness of observations is a domain-dependent factor. Yet, the constraints on the availability of information is a key factor in the coordination. This is a primary motivation for analyzing communication as a means of information transfer when observations alone are not adequate.

6.1.4 Novel States

As a final consideration, while the bulk of work in coordinating with an unknown teammate has assumed the possession of defined teammate types or relevant past experience, little attention has been given to scenarios where the agent lacks significant prior knowledge. This is a plausible concern in two immediate cases. When an agent is first operating in a domain, if it has not been given human-authored teammate models, it may have difficulty identifying even the most common of team strategies. Such scenarios are related to the problem of *exploration* in reinforcement learning, wherein an agent must gather information before it may accurately model the population of teammate strategies within a

domain.

Secondly, whether it be through rare stochastic transitions or through inaccurate teammate predictions, the actions of the team may transition the world state to one that has not been encountered before, reducing the effectiveness of leveraging past experience. Barrett et al. [12] touch on this topic briefly when discussing the ability of their approach to accurately predict team behavior after relatively few observations. The authors note that due to the static initial conditions as well as the consistency in behavior across teams, only a relatively small number of states are reached. Violations of such domain properties can leave the team in novel areas of the state space, in which one or more agent has little experience coordinating as a team. In such circumstances, it is necessary to either select a "safe" strategy [31] or use an alternate mechanism to observation-based approaches for coordination.

6.2 Coordination through Sharing Information

At its heart, the uncertainty between agents in a multiagent system is a result of the asymmetry of information. This can be information regarding the state of the world in which a task is being completed, the current nature of the task itself, and the individual tasks or plans of the other agents operating with the shared space. Coordination specifically requires a degree of shared information such that the agents can accurately model and plan around the uncertainties in a cohesive manner. This is the underlying concept of shared mental models [85, 118, 98]. Once agents have distributed the necessary information, it is possible to reach consensus on a joint plan of action, often an interleaving of actions from individual plans. On the contrary, if a vital piece of information is not shared among agents, individuals may improperly account for the intended actions of

their peers, resulting in unexpected outcomes. Here, we discuss the concepts of shared information and intentions in teamwork, motivating communication as a means of information-gathering in ad hoc team settings.

6.2.1 Intentions

The theory of intentions for a single agent was presented as an extension of the belief-desire model, having failed to describe the whole of rational behavior [27]. Intentions can be used to describe the commitment of an agent to a course of action, as a product of its beliefs and desires. Yet, in doing so, we must consider their transient nature, as intentions are subject to change, such as when an agent believes the intention has been fulfilled or has become impossible to fulfill. The act of planning, then, can be taken as the formation of and commitment to intentions, as guided by the particular beliefs an agent has regarding the world as well as its current desires. Cohen and Levesque [37] discuss this line of reasoning applied to agents without the specification of a particular agent architecture, as the terms broadly describe the execution of any rational agent. If goals (alternatively, *desires* in belief-desire-intention models [49]) motivate an agent's to act and beliefs constitute the requisite knowledge an agent possesses, intentions represent the course of action, grounded in beliefs, toward achieving such ends.

Under such characterization, intentions are inherently individual, based in independent perspectives and experiences. Coordination, then, requires some degree of alignment in intentions, one which motivates communication as a necessary consequence [38]. Cohen and Levesque [79, 38] note that individual perspectives are prone to cases of confusion when one or more agents diverge from a mutual plan. With such an observation, an agent may attempt to reason about why another diverged, assigning a cause and updating its beliefs. However, the possible reasons

for such deviation—e.g. hidden information, a failed attempt, conclusion that the plan is impossible—may cause the observing agent to reach an incorrect conclusion. Furthermore, in the initial formation of individual intentions toward a joint effort, conflicting beliefs of the agents can result in correspondingly conflicting intentions. To prevent this outcome, joint intentions were proposed [68, 38].

In essence, a joint intention requires the mutual belief that all members intend for the collective action to occur and, furthermore, that the team members retain mutual belief of that act while coordinating. If the goal is accomplished, considered impossible, or deemed irrelevant, one or more agents may adopt new goals and, consequently, new intentions individually. However, under the framework of joint intentions, the agents must establish both the mutual belief of the state of the old pursuit (succeeded, failed, or otherwise) and facilitate the adoption of a new joint intention by all members [68].

In contrast, for SharedPlans, Grosz and Kraus [57] do not strictly require joint intention or an exit protocol in the event that a subtask of a larger goal fails. Individual intentions are shown to be sufficient in collaborative action, though mutual belief of the task and its possible completion remain as an initial requirement. In the event an agent observes a part of the plan fail, it is left up to the agent as to whether communication to the other members is required for continual pursuit of a goal. If the agent can adjust its participation in the joint action such that the original goal can still be realized without requiring the alteration of the other agents' beliefs and intentions, communication is unnecessary.

The discussion of intentions—individual or shared—and mutual information is of interest to the ad hoc community, due to the uncertainty of a teammate's role in the team as well as its commitment to a particular course of action. A common assumption in ad hoc domains is that of non-recursive agent teammates

[46, 14, 10], i.e. teammates that do not model or reason over the intentions of the ad hoc agent. This results in rather asymmetric beliefs and intentions, as the ad hoc agent may consider its plan within the context of the team. The ad hoc agent's attempt at coordination, then, is not strictly to pursue the given goal but potentially to manipulate the individual courses of the teammates for the pursuit of the mutual goal. Moreover, by definition, the team cannot hold a joint intention or mutual belief, potentially resulting in the divergence of mental states, as the non-recursive teammates are not motivated or required to inform the ad hoc agent when their intentions change, as occurs in the case of changing teammate strategies discussed in Chapter 5. The apparent conflict between the ad hoc assumption—agents possessing minimal information regarding the decision processes of their teammates—and necessity of mutual beliefs in other coordination frameworks (SharedPlans, joint intentions) remains an open topic of discussion.

6.2.2 Shared Mental Models

During their description of teamwork, Cohen and Levesque [38] mentioned the concept of a shared mental state, "the glue that binds teammates together." This concept is mirrored across related team research as *shared mental models* [85, 118, 98]. A shared mental model represents a collection of joint information regarding the world state, expected transitions, potential tasks, knowledge among teammates, and the behavior or roles of the team members. The theory, stemming from work in psychology, suggests that team members who possess information regarding both the task and the role of each member can better anticipate the needs and requisite actions of their collaborators. This shared model of the process permits the team to coordinate effectively, often with reduced communication [98].

In recent years, the concept of shared mental models has been applied to

multiagent decision systems. Yen et al. [152, 153] proposed CAST—Collaborative Agents for Simulating Teamwork—as a model for teamwork among distributed, heterogeneous agents. The CAST architecture makes decisions according to the interplay of the individual and shared mental models, proactively communicating when the expected utility of sharing information exceeds that of not sharing. The explicit concept of a shared mental model was later formalized for use by agent systems [71].

6.2.3 Assigning a Value to Information

Similar to the belief-desire-intention model for multiagent interaction, the concept of shared mental models is often implicitly present in a system. In the communicative multiagent team decision problem (COM-MTDP) [108], the empirical evaluation featured an escort agent tasked with destroying an enemy radar so that a transport agent could safely pass to the goal. If the transport is too far from the radar when it is destroyed, it does not observe the event and proceeds cautiously. The escort must consider the possibility of communicating the radar's destruction, introducing the knowledge to the teammate by bringing about a mutual belief. Likewise, in the multi-armed bandit scenario [136, 10], teaching agents must consider the trade-off between introducing more information to the team via demonstration or allowing teammates to take potentially sub-optimal actions.

Such reasoning over the *potential* but perhaps uncertain benefit of exchanging information is posed as a decision-theoretic task [55, 53, 51, 108, 10, 152]. Due to the assumption that communication has associated resource costs, agents must often consider the potential result of a communicative act and judge whether the benefit exceeds the cost, assessing what is called the *value of information* [63], a topic we will return to in Chapter 7. Gmytrasiewicz et al. [55] illustrated this

concept in team domains with two types of messages, those of intentions and those containing world information. Further work extended the types of communication to questions, proposals and threats, imperatives, and statements of knowledge and beliefs [53]. Naturally, the capability to reason over the exchange of information brings as an additional consideration the potential for misinformation, should a teammate supply intentionally misleading or incorrect information [51].

6.2.4 Communicating State Information

In partially observable environments, sharing information allows agents to establish mutual beliefs. Due to the homogeneity of agents in many MAS applications, it is common for agents to assume perfect knowledge about the decision processes of their teammates with the sole exception being the current state of beliefs regarding the world state [108, 116]. As the set of possible observations and their associated probabilities is mutually known, the agents can broadcast their individual observation histories, then perform belief revision identically to update their teammate models as well as sync each agents' current estimation of the world state.

In domains of asymmetric information gathering, an agents may query information from others [117] for its own planning purposes or share information proactively, influencing the behaviors of their collaborators [136, 10]. For example, teaching agents in the multi-armed bandit scenario in [136, 10] possess perfect information and reason about what information to communicate to agents learning the partially-observable payout distributions of the bandits. Decisions to share information, then, depend on the factors of timing in content, corresponding to the two questions *When should an agent communicate?* and *What should an agent communicate?* posed by Maayan Roth et al. [116, 115].

6.2.5 Communicating Plans

A key element of planning in a multiagent domain is the accurate projection of the actions of other agents. While this can be learned over time [14, 13], it is often more direct to communicate regarding the expected behavior of coordinating agents. Stone and Veloso [137] discuss the communication of roles, setplays, and formations for coordination in robotic soccer. Tan [144] demonstrated the enhanced performance of agent teams when learned policies are shared.

SharedPlans [57] provides the most well-specified theory of communication and negotiation of plans in a multiagent setting. SharedPlans are constructed as hierarchical abstractions of actions required to complete a task. High level task are divided into lower level task, which are eventually decomposed into primitive actions. A task is assigned to an agent or group of agents if the team mutually believes the assigned party can complete the task. Through the communicated assignment of agents to single and multiagent subtasks, a full plan is realized.

6.3 Active Learning and Inference

With the observation that most ad hoc teamwork research primarily focuses on learning models of behavior or inferring teammate policies, we are motivated to seek out uses of communication that aid in the learning process, whether it be through informing a model directly or providing support to a model which does not have sufficient information to make accurate predictions.

In contrast to learning approaches that attempt to construct a model over a given set of labeled examples (*supervised* learning) or a sequence of incoming data points (*online* learning), active learning considers an approach in which the learning algorithm decides which of a set of unlabeled data points should be labeled

and incorporated into the model [125]. This process is motivated by the observation that certain data points may have a larger impact on the model once labeled, allowing a more accurate model to be trained from a comparatively small portion of the data. In tasks where labeling incurs a cost, it is beneficial to strategically decide which points are selected for labeling, minimizing the overall cost while maximizing the inferential or predictive power of the model. However, the exact computation of a set of optimal points to label is intractable for large problems due to the exponential space of subsets to evaluate. Consequently, many approaches select data to be labeled according to proxy measures, such as selecting data points with maximal uncertainty under the current model or points which are expected to most reduce the model's variance [125]. Inspired by this direction, we introduce similarly motivated heuristics in Chapter 10, though designed specifically for ad hoc team settings.

Judah et al. [72] applied active learning to imitation learning problems for MDPs, where a learning agent attempts to imitate a teacher's policy. The goal of the agent is to desired policy it should follow, and it is allowed to query the teacher for the action that should be taken at a given state. At a high level, this application of active learning shares many qualities with that of this work, specifically an agent trying to learn a policy known by another, with communication as the mechanism for acquiring necessary information.

In a similar vein, active inference considers the problem of obtaining information for the process of inferring a result given a trained model [22, 111]. This research direction has not received much attention and primarily has been used for collective classification tasks, in which data are represented in a graph structure, the links of which indicate a form of relationship between the data which may be used to infer labels of yet unlabeled data. For example, movies that

share a director are naturally related and may hold key information for use by recommendation systems [22]. Again, while the ad hoc teamwork problem has unique properties that set it apart from these works, we are nonetheless similarly motivated in acquiring information to improve inference of teammate policies.

6.4 Summary

We have provided an overview of conditions under which purely observation-based approaches can leave an ad hoc agent uncertain as to the policies of its teammates. These include but are not limited to observing novel teammate policies, coordinating in time-constrained scenarios with few observations, observing non-discriminative behaviors, and collaborating in states where little prior experience can be leveraged. As teammate policy uncertainty is at the root of each of these conditions, we propose communication as a mechanism for acquiring vital information for the coordinated effort. To this end, the remainder of this document will consider the role of communication with the purpose of resolving behavioral uncertainty, employing information- and decision-theoretic reasoning to elicit the minimal amount of policy information necessary for an ad hoc agent to successfully coordinate with unknown teammates.

Chapter 7

Communicating Policies

We are interested in analyzing the exchange of policy information via explicit communication as motivated in the previous chapter. In an effort to do so, we characterize the position of the uncertain agent, its policy, as well as the set of information it possesses as an MDP over the space of potential policies or strategies a teammate may be pursuing. Constructing this problem as an MDP allows us to describe many of the properties and challenges of the problem as well as to provide a formal model under which an agent may compute optimal or approximately-optimal communication policies.

Across many communicative multiagent frameworks, such as the COM-MTDP model [109] and STEAM [141], communication is constrained to the sharing of observations or direct state information [10, 117]. As the policies of teammates are the source of uncertainty in ad hoc teamwork, it follows that policy information is a promising target for communicative acts. Sharing information between agents aids in establishing a shared set of beliefs, reducing the likelihood that agents diverge in their beliefs with respect to the joint plan. Analogously, the reduction in uncertainty regarding a teammate's policy consequently permits a coordinating agent's ability to align its own policy to that of its peers.

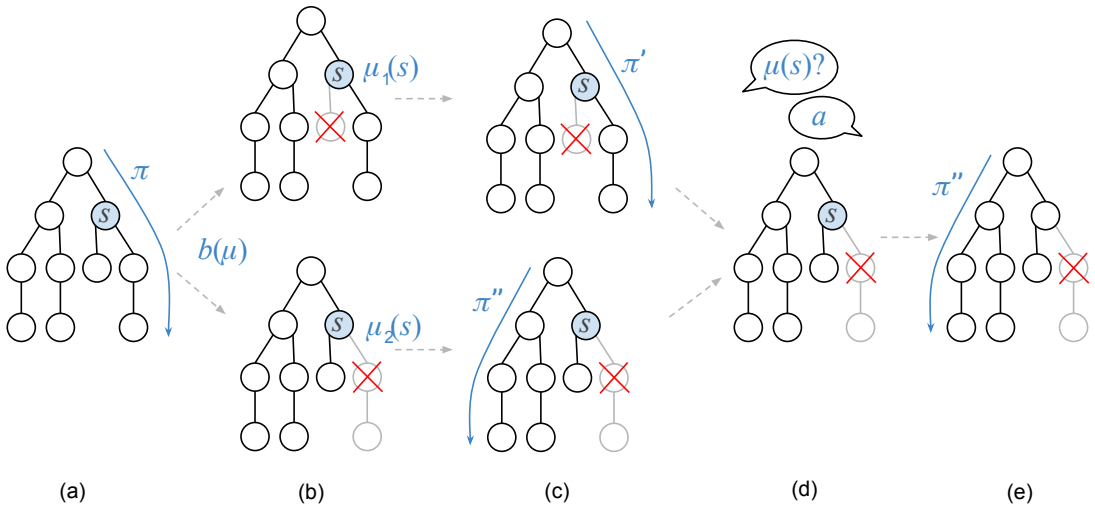

Figure 7.1: Communication analysis for revising plans. (a) Given a finite-horizon policy graph, the communication process identifies a point of uncertainty, $\mu(s)$, (b) considers potential outcomes from a teammate's decision, as predicted by $b(\mu)$, (c) recomputes the agent policy π for each case, and (d) initiates communication if the potential change in expected policy value exceeds the cost of communication, (e) resulting in a revised policy, π'.

As policies are naturally divided into discrete elements—state-action assignments—we propose using these individual decision points as targets for exchanging information regarding teammate policies[1]. From these state-action pairs, two types of communicative acts are immediately obvious: instructive commands, instructing a teammate to perform a specific action when a state is encountered, and policy queries, requesting what action the teammate intends to perform at the given state. As we are primarily interested in inferring the policies of teammates, we will focus on policy queries as an application of active inference in ad hoc team domains. Figure 7.1 depicts a high-level overview of the communication process in which an agent reasons about the value of querying a teammate for policy information subject to a cost associated with the communicative act. This chapter

[1]In many cases, communicating higher level coordination information, such as roles, goals, or subtasks, may be more appropriate. However, such content is domain-specific, and we are primarily motivated by a domain-independent perspective of ad hoc teamwork.

defines the process in detail.

7.1 Assumptions

In order to address the problem of policy communication in ad hoc teams, we make the following assumptions regarding the capabilities of the team members:

- The team is coordinating on a task with finite, discrete state and action spaces.
- Team members share a state and action representation that can be communicated.
- All team members are capable of communication, though the exact form of which is left to the domain or application.
- The team members other than the ad hoc coordinating agent follow stationary policies. We leave extensions to non-stationary policies for future work (Section 12.3.2).
- All team members answer policy queries honestly.

Each of these assumptions can, in practice, be altered, but such an application may necessitate considerations not covered in this research. For example, the assumption of a finite, discrete state space naturally forms a set of queryable policy information; in contrast, under an infinite, possibly continuous state space, it is necessary to sample or otherwise construct a finite set of policy information from which to query.

7.2 States of Information

For the purpose of providing a clear vision of how communicating policies leverages and improves information obtained from observations of teammate behavior, we must outline what is meant by an *information state* in this context. Here, we use the term *information* as a notion of some observation or obtained piece of data that potentially impacts the uncertainty an agent has with respect to its situation. With respect to POMDPs, information is embodied by the observations an agent receives, which allows the revision of a belief distribution over partially-observable states. In general, an information state is a collection of discrete pieces of information obtained up to a point in time, t. For our application, we define an information state, I_t to be the collection of partial policy information gathered from observations as well as from communicated policy information. For the purely observation-based perspective of the ad hoc teamwork problem, as covered in Chapter 2, an information state can be defined simply by the observation history, $I_t = \mathsf{O}_t = \{o_1, o_2, \ldots, o_t\} = \{(s_1, \bigtimes_{i=1}^{n} a_{i,1}), (s_2, \bigtimes_{i=1}^{n} a_{i,2}), \ldots, (s_t, \bigtimes_{i=1}^{n} a_{i,t})\}$. Without any loss of generality, we will focus on the observations and information state specific to a single teammate, such that $\mathsf{O}_t = \{o_1, o_2, \ldots, o_t\} = \{(s_1, a_1), (s_2, a_2), \ldots, (s_t, a_t)\}$.

For the communicative case, the information state further incorporates information gathered through policy exchanges, i.e. $I_t = \mathsf{O}_t \cup \mathcal{Q}_t$, where $\mathcal{Q}_t$ is the set of policy information queried among a subset of teammate policies. Here we distinguish the contributions of observations and queries, which may occur in varying quantities. For example, while the agent receives an observation per step, resulting in t observations at t actions, we allow for more flexible quantities of policy information queries, as we will explain in further detail in Section 7.5.

Given an information state, I_t, $|I_t| = m$, and a belief prior, b, we can compute

the resulting belief, b', by the following:

$$\begin{aligned} b'(\mu) &= \Pr\left(I_t \mid \mu\right) b(\mu) \\ &= \Pr\left((s_1, a_1) \cap (s_2, a_2) \cap \cdots \cap (s_m, a_m) \mid \mu\right) b(\mu) \\ &= \prod_{k=1}^{m} \Pr\left((s_k, a_k) \;\middle|\; \bigcap_{j=1}^{k-1} (s_j, a_j), \mu\right) b(\mu) \qquad \text{(Chain Rule)} \end{aligned}$$

This amounts to iteratively applying the belief revision function, $\mathcal{B}$, incorporating each observation and communicated policy action. As the fully observable world state remains unchanged, the revised belief state becomes $\mathsf{b}' = \left(s, \bigtimes_{i=1}^{n} b'_i\right)$ where s is the current state. In this way, we use information states to generalize accumulated information from varied sources but for the unified purpose of revising beliefs. As each information state is associated with a belief state[2], we may begin to analyze how changes in the information state correspond to changes in the expected utility of an agent's individual policy under the remaining uncertainty over teammate policies. Recall from Chapter 2 the value function over beliefs, $V(\mathsf{b})$. Given the correspondence between information states and belief states, we can similarly express the value function in terms of information state, i.e. $V(I)$. For the purely observation-based version, the value equation formulation follows directly from $V(\mathsf{b})$. However, under the possibility of querying further policy information directly, we must consider multiple new aspects of the problem, specifically

- How does querying policy information affect the agent's expected reward?
- Similarly, how does the new information affect the agent's individual policy?
- How does an agent select which policy information to query from a team-

[2]With this connection between information states and belief states establish, we note that from here forward, we will use $\mathsf{Vol}(I) = \mathsf{Vol}(\mathsf{b}')$ interchangeably.

mate?

7.3 The Value of Information

The utility of a set of new information impacts the expected value of an agent's individual policy in two manners. First, it reduces uncertainty over the predictions of teammate actions, reducing the variance of possible outcomes. Consider an agent with current information I, policy π, and expected value $V_\pi(\mathsf{b})$. By obtaining new information, the agent moves to a new information state, I', and a new corresponding expected value, $V_\pi(\mathsf{b}')$. This change does not always improve the expected value of an agent's current policy, however; in the case that the agent is overestimating the likelihood of an optimistic outcome, discovering the discrepancy can decrease the projected cumulative reward.

Secondly, new information allows the agent an opportunity to change its policy. There are two broad conditions under which this occurs: when the value of the agent's current policy trajectory drops below that of an alternative policy or when the expected value of an alternative trajectory raises above that of the current trajectory. In either case, as $V_{\pi'}(\mathsf{b}') > V_\pi(\mathsf{b}')$, the agent is incentivized to alter its policy in pursuit of maximizing its expected payoff. This gain in expected utility is called the *value of information*, and is given by

$$\mathsf{Vol}(\mathsf{b}') = \max_{\pi'} V_{\pi'}(\mathsf{b}') - V_\pi(\mathsf{b}') \tag{7.1}$$

In estimating this value before the information is received, it is useful to conceptualize the *expected value of information*,

$$\mathbb{E}_{\mu(s)}\left[\mathsf{Vol}(\mathsf{b}')\right] = \mathbb{E}_{\mu(s)}\left[\max_{\pi'} V_{\pi'}(\mathsf{b}') - V_\pi(\mathsf{b}')\right] = \mathbb{E}_{\mu(s)}\left[\max_{\pi'} V_{\pi'}(\mathsf{b}')\right] - V_\pi(\mathsf{b}). \tag{7.2}$$

This expectation allows us to reason about the expected gain of utility across each of the potential responses to a policy query, as weighted by the relative likelihood of each. Furthermore, it allows us to reason about the potential value of new information before knowing precisely what the response will be. As the objective of an agent selecting a coordinating policy is to maximize the expected reward, the act of soliciting further information is an opportunity to refine a coordination policy toward this end.

7.4 The Policy Communication Decision Problem

We formalize the problem of communicating policy information as a decision problem across information states with queries as potential actions. We pose this decision problem as a Markov decision process (MDP) such that we can employ existing solution techniques and reason about methods of approximating optimal communication policies. The next few sections establish the components of the problem.

As we alluded to before, we are motivated by the need to resolve uncertainty for one or more teammates' policies. It is not as simple as evaluating all subsets of a policy we could query, however, as the response may influence the which parts of the teammate policy should be resolved next. This process, then, is interactive, forming a graph of question-response events, of which any query begets one of a set of possible responses, and the result of which conditions future queries.

The *Policy Communication Decision Problem* (COMMDP) is represented as an MDP of the form $\langle \mathcal{I}, \mathcal{Q}, \mathcal{M}, \mathcal{U}, \lambda \rangle$, wherein $\mathcal{I}$ is the set of information states, $\mathcal{Q}$ is the set of possible policy queries, $R : \mathcal{I} \times \mathcal{Q} \times \mathcal{I} \longmapsto \mathbb{R}$ is the value of transitioning

between information states given a query and response, $M : \mathcal{I} \times \mathcal{Q} \times \mathcal{I} \longmapsto [0, 1]$ is the likelihood function of an information state transition given a query, and $\lambda \in [0, 1]$[3] is the discount factor for the COMMDP. For this work, we will use a discount factor of $\lambda = 1$ and omit it from the analysis for clarity and conciseness.

7.4.1 Information States

An information state, $I \in \mathcal{I}$, represents a specific set of knowledge regarding the policies of an agent's teammates. Formally,

$$\mathcal{I} = \bigcup_{P \in \mathscr{P}(S \times \{1,\ldots,n\})} A^P \cup \{I_{\text{stop}}\}$$

is the space of information states, where $\mathscr{P}(X)$ is the power set of X and $\{I_{\text{stop}}\}$ is a terminal information state. Elements P are collections of state and agent index pairs, with A^P being mappings of the pairs to actions. Elements $I \in \mathcal{I} \setminus \{I_{\text{stop}}\}$, then, are subsets of agent policy information known, as outlined earlier in the chapter. In practice, the information states reached by a coordinating agent are relatively sparse in coverage of the underlying policy space, particularly in large, heavily-branching state spaces. Given a finite state and action space, the number of unique information states is on the order of $(|A| + 1)^{n|S|}$ where n is the number of teammates with which the agent is coordinating.

7.4.2 Query Actions

Each query action resolves a single state-action pair of a single teammate's policy. Furthermore, for this work, we assume both that the teammate policies

[3] As the COMMDP is defined over a finite information state space, in which agents may obtain but never lose information, communication is guaranteed to terminate, forming a finite-horizon decision problem. This observation permits $\lambda = 1$.

are stationary and that the teammates relay accurate policy information; we leave extensions of non-stationary policies and noisy communication for future work. Here, $\mathcal{Q} = (S \times \{1, \ldots, n\}) \cup \{q_{\text{stop}}\}$ is the set of policy queries to an agent, where q_{stop} is an action halting the line of queries, transitioning to I_{stop}.

7.4.3 Information State Transitions

Once queried, a teammate will respond with the appropriate policy information. From the ad hoc agent's perspective, the possible results are stochastic, with probabilities of each potential policy action given by

$$\mathcal{M}\left(I, q, I'\right) = \mathcal{M}\left(\left(s_0, \bigtimes_{i=1}^{n} b_i\right), (s_t, j), \left(s_0, \bigtimes_{i=1}^{n} b'_i\right)\right)$$

$$= \begin{cases} 0 & \text{if } I = I_{\text{stop}} \\ 1 & \text{if } q = q_{\text{stop}} \text{ and } I' = I_{\text{stop}} \\ 1 & \text{if } I = I' \text{ and } \exists! a \ (s_t, j, a) \in I \\ \sum_{\mu} \mathbb{1}_{\mu_j}((s_t, a)) b_j(\mu_j) & \text{if } |I' \setminus I| = 1 \text{ and } I' \setminus I = \{(s_t, j, a)\} \\ 0 & \text{otherwise} \end{cases}$$

7.4.4 The Reward of Communicating

Given an information state I_t, a query action q, and its resulting information state I_{t+1}, we can calculate the payoff from communicating as the difference in expected utility of the corresponding policies associated with each information state. In many domains, it is assumed that the act of communicating is associated with a fixed cost, reflecting the resources—such as time or energy—required to exchange information. In these instances, the payoff function can additionally incorporate the cost, $C : \mathcal{Q} \longmapsto \mathbb{R}_{\geq 0}$. Therefore we formulate the reward function

as

$$\mathcal{U}(I, q, I') = \begin{cases} \mathsf{Vol}(I') - \mathsf{Vol}(I) - C(q) & \text{if } q \neq q_{\text{stop}} \\ 0 & \text{otherwise} \end{cases}$$

It is important to note a distinction between this view of evaluating communication and that described in Section 7.3. In the earlier formulation of the value of a communicative act, the evaluation implicitly considers each act in isolation. This approach is correct in instances where there is no added benefit to initiating more than a single act of communication or where agents are restricted to communicating only a single item. As our domain necessitates the exchange of sets of information, similar to that in [116], we consider as the reward the change in the value of information for the set of information in each information state. In other words, the reward function is the change in the value of information from adding a query-response pair to an information state less the cost of the query.

7.4.5 Termination Criteria

The precise portion of a policy necessary to clarify is not of a fixed size but is dependent on the current uncertainty, risk involved across future horizons, and also what information the agent receives as it begins its line of queries. Unlike other multiagent communication decision problems [106, 116], we do not limit the agent to a single exchange of information at each step of coordination but rather allow the decision process to decide whether or not further information should be queried. This need is in part due to the interactivity of the process, as the need for further information is conditioned on the queried teammate's response, which is inherently unpredictable. Naturally, this raises the question: *When should the communication process terminate?* It is obvious that in finite-horizon coordination

domains, only $|S|$ policy states can be queried per teammate. In practice this never occurs, as state spaces of the domain are typically too large to exhaustively query. Alternatively, the value of information crucially yields some ending criteria for the communication process.

Recall that the set of query actions contains an action for q_{stop}. When this action is the optimal communication policy action, the communication exchange terminates. As this action has a cost, $C(q_{\text{stop}}) = 0$, it will only be the optimal communication policy action when all alternative lines of queries each have a negative cumulative expected value, i.e. when the cost of communication exceeds the value of the information exchanged.

Alternatively, we can establish an upper bound for the maximum cumulative expected value of information. When this upper bound is less than the cost of a query, it is clear that no line of querying could exceed the incurred cost. We discuss an initial bound for this value in Section 8.2 as well as how cost constrains the maximum quantity of queries in Section 11.2.4 .

7.5 Planning, Communicating, and Acting

With a method in place for eliciting necessary policy information from teammates, an ad hoc agent can now utilize all possible sources of information while attempting coordination. The model of a communicating ad hoc agent is depicted in Figure 7.2. At the beginning of an instance of coordination, an agent possesses a distribution over potential teammate policies, either constructed from past experience or generated by another process. In the first stage, the agent computes a policy maximizing its expected payoff over some finite horizon. With a policy in hand, it considers the communication problem, as we have just outlined. It computes a policy in the information space, then executes a series of

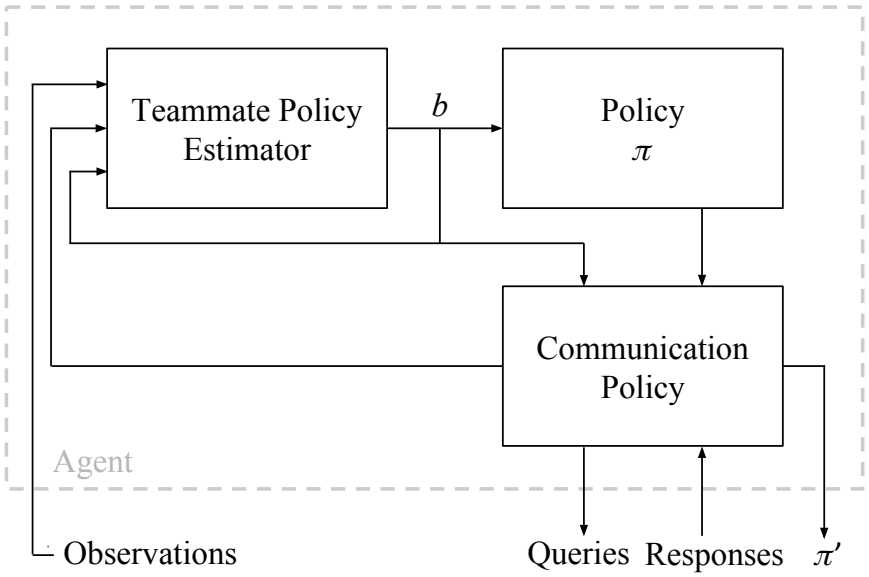

Figure 7.2: A model of an ad hoc agent capable of observing and soliciting policy information from teammates.

communicative acts between it and its teammates, querying for policy information, receiving responses, and updating its internal information state. Once the communicative policy specifies the q_{stop} action, the agent ceases communication, updates its domain-level policy ($\pi \rightarrow \pi'$) by revising its beliefs from the gathered information from the communicative process, and selects its next policy action accordingly. Once the team's joint action is performed, the agent records the observation, updates the model of each of its teammates via the calculating a posterior distribution over possible policies, and repeats the process.

Chapter 8

Theoretical Characterization

Computing a communication policy over an information state space that is exponentially larger than the underlying domain state space is clearly intractable for problems of sufficient scale. It may be, however, that properties of its specific construction may be exploited to provide more computationally feasible, optimal solutions or near-optimal, heuristic techniques. In this chapter, we examine various properties of the communication decision problem in order to better understand how we may design algorithms for overcoming its computational requirements.

8.1 The Information State Space

The information state space underlying the communicative decision process, as we have constructed it, possesses a rather orderly structure. For clarity, we will illustrate the size and shape of the state space given an initial information state, I. This state encodes policy information for some k state-action pairs out of a maximum set of $n|S|$ policy actions for n teammates. Consider the space of queries the coordinating could send to its teammates. First, we observe that querying policy information that is already known cannot improve the agent's policy. Given this

observation, we remove from consideration any query of policy information from a teammate that is already covered by the information state, leaving only queries that grow the information state. There are, then, $n|S| - |I|$ queries available. Each query can result in one of $|A|$ responses, yielding $|A|\left(n|S| - |I|\right)$ new information states, each corresponding to a unique partial view of the team's overall joint policy. Due to this additive nature of the information state constructions, the state space takes the form of a directed acyclic graph.

It is useful to consider the information states reachable given a fixed number of queries. We will refer to these sets of information states as *horizons* of information, e.g. the first horizon represents the set of information states reachable given an initial information state I and all possible query response pairs for a single query. Clearly, due to the finite space of potential queries, there are at maximum $h^* = n|S| - |I|$ horizons. Furthermore, each horizon, h, contains $\binom{h^*}{h}|A|^h$ information states, totalling $\sum_{h=0}^{h^*}\binom{h^*}{h}|A|^h = (|A|+1)^{h^*}$ information states across all horizons. This state space represents all of the partial information sets the agent could possess through querying from its current information state. Notably, this space is exponential in both the size of the team and the size of the domain state space.

Similarly, we can enumerate all possible trajectories from a given information state, I, to states of complete information, i.e. information states include the entire policies of each teammate. At each horizon, h, there are $h^* - h$ queries available per information state, each with $|A|$ potential responses. The number of unique query trajectories is, then, $\prod_{h=0}^{h^*-1}\left(h^* - h\right)|A| = (h^*)!|A|^{h^*}$ in total. Fortunately, as we are operating on unordered sets of information, trajectories various orderings of the same set of query-response pairs result in identical information states, allowing us to employ dynamic programming to avoid the evaluation of trajectories individually but rather focus computational effort on coverage of the

information state space.

8.1.1 A Note on Submodularity

We observe that the value of information is a set function over subsets of policy information, as established in our construction of information states. The decision problem of finding an optimal communication policy is simply the maximization of the set function subject to a cost per item of the set. In many respects, this mirrors the process of what is called *submodular maximization* [94], a well-studied class of maximization problems wherein the goal is to find a subset $S \subseteq N$ maximizing $z(S)$ where z is a submodular function. Submodularity is related to the concept of diminishing returns, and a function is considered submodular if for $S \subseteq T \subseteq N$, $z(S \cup \{x\}) - z(S) \geq z(T \cup \{x\}) - z(T) \quad \forall\ S, T \subseteq N,\ \forall x \in N \setminus T$. In other words, the marginal benefit of adding $\{x\}$ to S is at least as much as adding it to T. Such maximization problems admit greedy approximation algorithms whose solutions are within a factor of $1 - 1/e$ of the optimal value when z is also monotone. Unfortunately, while the expected value of information is monotone, it is not submodular, which we will illustrate by counterexample.

Consider a simple scenario of three states, $\{s_1, s_2, s_3\}$, and two agents, $\{\alpha_1, \alpha_2\}$. In s_1, α_1 may choose to end the scenario with an immediate payoff of 8 or begin a game of coordination with α_2. If the agent attempts the coordination game, the game transitions to s_2. In s_2, both agents must attempt to coordinate their individual actions, $a_i \in \{a_1, a_2\}$, such that both agents select the same action, e.g. $\bar{a} = \langle a_2, a_2 \rangle$. If this occurs, the game transitions to s_3; otherwise, the game ends with 0 payout. For s_3, the agents must again attempt to coordinate their actions, $a_i \in \{a_1, a_2\}$. However, upon a success, the game terminates with a payout of 10. Furthermore, α_1 models α_2 with a uniform distribution over the four possible

successful joint policies that achieve the maximum payout. As we can see in Table 8.1, it is clear that adding a second state-action pair to the information state in this example is of greater value than adding the first in every case, e.g. $\mathsf{Vol}(\{(s_2, a_1)\} \cup \{(s_3, a_1)\} - \mathsf{Vol}(\{(s_2, a_1)\}) > \mathsf{Vol}(\{\} \cup \{(s_3, a_1)\}) - \mathsf{Vol}(\{\})$.

I	$V_{\pi'}(s_1)$	$V_{\pi}(s_1)$	$\mathsf{Vol}(I)$
$\{\}$	8	8	0
$\{(s_2, a_1)\}$	8	8	0
$\{(s_2, a_2)\}$	8	8	0
$\{(s_3, a_1)\}$	8	8	0
$\{(s_3, a_2)\}$	8	8	0
$\{(s_2, a_1), (s_3, a_1)\}$	10	8	2
$\{(s_2, a_1), (s_3, a_2)\}$	10	8	2
$\{(s_2, a_2), (s_3, a_1)\}$	10	8	2
$\{(s_2, a_2), (s_3, a_2)\}$	10	8	2

Table 8.1: Value of information for each information state.

The consequence of non-submodularity is in the difficulty of evaluating intermediate information states between the initial information state and some future state where enough information has been gathered to improve the agent's policy. In this example, it is only upon querying the entire policy of α_2 that α_1 adapts its policy, yet the intermediate steps of querying only a subset of the policy of α_2 provide no value in isolation. As the required set of information could be arbitrarily large in a given coordination domain, the corresponding search through the information state space may be one of sparse rewards over large horizons.

8.2 Bounds on the Value of Information

For the purpose of exploring informed solution approaches to the problem of communication, we are motivated to establish bounds on the value of information and, correspondingly, the expected value of information.

Given a an agent with current information state, I, consider a new information state I' such that $I \subseteq I'$. We note that I' induces a new belief state, b'. By definition, the value of information is calculated as $\mathsf{VoI}(\mathsf{b}') = \max_{\pi'} V_{\pi'}(\mathsf{b}') - V_{\pi}(\mathsf{b}')$. Observing that $\max_{\pi'} V_{\pi'}(\mathsf{b}') \geq V_{\pi}(\mathsf{b}')\ \forall \pi$, we can see that the value of information and, consequently, the expected value of information must be non-negative.

The computation of an upper bound for the value of information is more involved. Consider the policy information contained in an information state I', again corresponding to b'. The known policy decisions form a set of constraints on the policy space of each teammate. Let $\bar{\pi}^* = \arg\max_{\bar{\pi}} V_{\bar{\pi}}(s)$ subject to $\mu_i(s_t) = a_i\ \forall (s_t, i, a_i) \in I'$. Then, $\bar{\pi}^*$ is the optimistic joint policy maximizing the MMDP under the constraints of the known policy information of the team. By construction, $V_{\bar{\pi}^*}(s) \geq V(\mathsf{b}')$ where $\mathsf{b}' = \left(s, \bigtimes_{i=1}^{n} b_i'\right)$. Therefore,

$$\mathsf{VoI}(\mathsf{b}') = \max_{\pi_0'} V_{\pi_0'}(\mathsf{b}') - V_{\pi_0}(\mathsf{b}) \leq V_{\bar{\pi}^*}(s) - V_{\pi_0}(\mathsf{b}), \quad \text{and}$$

$$\mathbb{E}\left[\mathsf{VoI}(\mathsf{b}')\right] = \mathbb{E}\left[\max_{\pi_0'} V_{\pi_0'}(\mathsf{b}')\right] - V_{\pi_0}(\mathsf{b}) \leq V_{\bar{\pi}^*}(s) - V_{\pi_0}(\mathsf{b}).$$

Note that in the event the agent's current belief state and policy match the value of the optimistic joint policy, the value of information becomes zero. This intuitively makes sense, as the beliefs hold that in every possible assignment of teammate policies, the agent's policy achieves the same degree of success as the optimistic result. This mirrors the case in SharedPlans [58] when a team member is assigned a task to complete, but the rest of the team is indifferent to how the task is completed. What matters is the mutual belief that the teammate *can* complete the task.

Furthermore, this upper bound can be used to form an admissible heuristic, though it tends to grossly overestimate the utility to be gained and, therefore, is

not particularly informative. In many cases, it is more practical to have an approximately accurate inadmissible heuristic than an inaccurate admissible heuristic, particularly as it may not be tractable to compute optimal communication policies in large domains.

8.3 The Impact of Cost on Communication

Under free communication, i.e. $C(q) = 0 \; \forall q \in \mathcal{Q}$, it is straightforward to show that the expected value of perfect information incentivizes communication policies which elicit the entire policies of all teammates. As the value is non-negative, it follows that the policy resulting from communicating the entire policies of all teammates must provide at least as much utility as any policy π constructed under partial information.

Typically, acts of communication are associated with a cost, such as time, energy, or some other exhaustible resource. For this work, we assume that each query has an associated cost, $C(q) \geq 0$, as it is rarely the case that an act of communication has an associated positive reward. Rather, the benefits of communication are from the implicit effect on the belief state and revised agent policy, a common observation in related literature [108].

The value of information is a measure of how much the agent's expected utility will change given new information. Earlier, we established an upper bound on this value, given by $V_{\bar{\pi}^*}(s) - V_{\pi_0}(\mathsf{b})$. As a rational agent chooses to communicate only when the expected utility of communicating exceeds the cost, we can use the upper bound to compute the maximum number of queries an agent has available before the total cost exceeds the potential benefit of further exchanges. Let $c_{\min} = \min_{q \in \mathcal{Q}} C(q)$. Then, $\lfloor c_{\min}^{-1} \left(V_{\bar{\pi}^*}(s) - V_{\pi_0}(\mathsf{b}) \right) \rfloor$ is the maximum number of queries before the total communication cost exceeds the maximum value to be gained.

In many practical applications, the cost of communication constrains the amount information that can be queried. In such scenarios, agents may not be able to prune the space of potential teammate policies to the precise policies being followed; rather, they will retain some degree of uncertainty during the act of coordination. However, uncertainty is not inherently problematic for coordination, as in certain circumstances, the agent may be indifferent to the policy choices of its team. This is particularly true where agents in a team work independently on individual tasks, which requires less direct coordination.

8.4 Summary

In this chapter, we have provided an initial characterization of the Policy Communication Decision Problem, describing the exponential size of the information state space, as well as the monotonic, finite accumulation of information that results from querying policy information. While the problem bears resemblance to submodular set maximization, the value of information is not submodular and, consequently, the communication decision problem currently provides no guarantees on greedy or approximate approaches. However, we are able to establish a bound on the total future expected value of information in a given information or belief state. Similarly, we can bound the maximum number of queries an agent can utilize before the cost exceeds the benefit of communication. Together, these two properties can inform communication policy algorithms, for example bounding the lookahead of a local search or providing criteria for early termination. Due to the complexity of this problem, we will focus primarily on approximate techniques for the remainder of this thesis.

Chapter 9

Greedy, Approximate Communication

The framework outlined in Chapter 7 outlines a decision problem of intractable size for any but the smallest of underlying coordination problems. We are motivated, then, to pursue approximate solutions that may substantially reduce the computational requirements of finding a communication policy. This chapter proposes and evaluates a initial, greedy approach to the policy communication problem.

9.1 Sidestepping Complexity

Recall from Chapter 7 the correspondence between information states in the communication decision problem and belief states in the coordination decision problem. Transitioning from one information state to another shifts the belief state, resulting in new predictions of teammate behavior as well as a potentially

The results of this chapter are presented in [120].

new individual policy for the coordinating agent. Practically, consider an agent whose coordination policy covers reachable belief states up to a finite horizon, h, forming a policy tree or policy graph the size of which is on the order of $|\bar{A}|^h$. With a new query-response added to the agent's information state, each belief distribution at each belief state in the policy graph would need to be revised. Furthermore, a complete solution to the communication decision problem would necessitate evaluation of many information state transitions, requiring complete reevaluation of the policy graph for each reachable information state[1]. Under such considerations, we identify the following targets for reducing computation complexity:

- Recomputing beliefs over an agent's policy.
- Recomputing the agent's policy, given new beliefs.
- Evaluating successive information state transitions to form a communication policy.

9.1.1 Approximating the Value of Information by Fixing Beliefs

Recomputing belief distributions for each belief state throughout an agent's finite-horizon policy effectively necessitates the recomputation of the entire policy. We observe that in many domains, differing agent policies beget transitions to correspondingly different regions of the underlying state space. For example, in the maze-based version of multiagent pursuit, an agent choosing one of two directions at a fork will rarely lead to the same state once the divergence is encountered. In such situations, the actions of an agent across the two choices—as represented by

[1]Section 8.1 details the size of the information state space and quantity of query trajectories.

two separate branches of a policy graph—are rarely correlated, and, consequently, knowing the policy of the agent in a state in one branch often does not impact the believed likelihoods of actions in another.

Furthermore, recall that successive beliefs in an agent's policy incorporate new information from the predicted observations. In our case, the observations are the very actions which are being queried by the agent to be incorporated into the agent's beliefs through communication. However, we treat these two methods of integrating new policy information identically. Therefore, belief states occurring after a state being queried already possess the policy information queried by an agent and, as a result, need not be updated by the communication process.

Given these two observations, we propose an approximate recalculation of an agent's expected utility under new information which only updates predicted teammate actions along trajectories from the agent's current belief state, b_0, and the belief state, b_t, corresponding to the queried state, s_t. The process for approximating the expected value of information a query using this approach is given in Algorithm 1.

9.1.2 Greedy Queries

While the true value of information is calculated over a series of successive queries, such a process is expensive to compute, particularly as the set of potential query actions at a given information state is on the order of the entire coordination state space. For this reason, we take a greedy approach to evaluating $\mathbb{E}[\mathsf{Vol}(I)]$, considering only a one-step lookahead return on querying each potential state in the teammate's policy. If $\max_{q \in \mathcal{Q}} \mathbb{E}\left[\mathsf{Vol}(I) \mid q\right] - C(q) > 0$, the agent queries its teammate and evaluates the next set of queries under revised beliefs. The success of a greedy strategy largely depends on the informativeness of individual queries,

Algorithm 1 Procedure for computing the approximate expected value of information of a given policy query to agent i for $\mu_i(s_t)$.

function ApproximateExpectedVoI(V, b_0, b_t, s_t, i)
    $U \leftarrow 0$
    for $a_i \in A_i$ **do**
        $V'(\mathsf{b}_0) \leftarrow \text{PropagateValue}(V, \mathsf{b}_0, \mathsf{b}_t, s_t, a_i)$
        $U \leftarrow U + \Pr(\mu_i(s_t) = a_i \mid \mathsf{b}_0)\ V'(\mathsf{b}_0)$
    end for
    return $U - V(\mathsf{b}_0)$
end function

function PropagateValue(V, b_0, b_t, s_t, a_i)
    $h \leftarrow t$
    $V' \leftarrow V$
    $\mathbb{B} \leftarrow \{\mathsf{b}_t\}$
    while $h \geq 0$ **do**
        $\mathbb{B}' \leftarrow \emptyset$
        for $\mathsf{b}_h \in \mathbb{B}$ **do**
            $\mathsf{b}'_h \leftarrow \mathcal{B}(\mathsf{b}_h, s_t, a_i)$
            $V'(\mathsf{b}_h) \leftarrow \max_{a_0 \in A_0} \mathbb{E}\left[\sum_{\mathsf{b}_{h+1}} \mathsf{T}(\mathsf{b}_h, \bar{a}, \mathsf{b}_{h+1})[R(\mathsf{b}_h, \bar{a}, \mathsf{b}_{h+1}) + \gamma V'(\mathsf{b}_{h+1})] \mid \mathsf{b}'_h\right]$
            $\mathbb{B}' \leftarrow \mathbb{B}' \cup \{\mathsf{b}_{h-1} \mid \exists \bar{a}\ \mathsf{T}(\mathsf{b}_{h-1}, \bar{a}, \mathsf{b}_h) > 0\}$
        end for
        $h \leftarrow h - 1$
        $\mathbb{B} \leftarrow \mathbb{B}'$
    end while
    return $V'(\mathsf{b}_0)$
end function

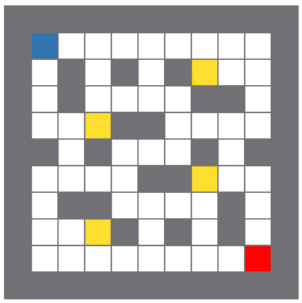

Figure 9.1: The maze used for the pursuit experiments. The coordinating agent is represented in blue, with the teammate in red. The fleeing evaders are represented by the four yellow cells.

as a myopic evaluation can fail to find *sets* of queries that hold greater value than the sum of the individual evaluations, as discussed in Section 8.1.1. We evaluate non-greedy communication in Chapter 11.

9.1.3 Empirical Evaluation

We test the greedy, communicating ad hoc agent in a maze shown in Figure 9.1, which depicts the initial configuration of the team and the fleeing evaders. While simple, 5.15×10^{10} unique placements of the agents and evaders 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.

9.1.4 The Coordinating Agent

We test the proposed greedy, approximate communication approach with two variants of a coordinating ad hoc agent. For decision-theoretic planning, the agent uses Upper Confidence Bounds for Trees (UCT) [75], a version of Monte-Carlo tree search (MCTS) [34] balancing exploration of novel policies and exploitation

of well-performing policies. Moreover, we implement two methods of modeling teammates, the first of which we will refer to as the *No Priors* agent, referring to a complete lack of experience coordinating with other agents in the domain. This model initially predicts the actions of its teammate uniformly; over time, the model predicts actions proportional to frequency of observations of the action, $n(a_i, s)$, as shown in Equation 9.1. We add a constant smoothing factor of 1 to ensure every action has non-zero probability. For the evaluation, we are interested in how the need and use of communication changes over repeated trials with a teammate, as the agent learns to predict the teammate's policy more accurately.

$$\Pr(a_i \mid s) \propto n(a_i, s) + 1 \tag{9.1}$$

Additionally, we demonstrate the approach with a second type of coordinating agent, which we will call *With Priors*, utilizing a set of known policies, one for predicting pursuit along the shortest path to each potential target. This approach mirrors that of existing work [14, 122] as well as the approach used in Chapter 5. The agent updates a belief distribution, initially uniform, over policies, according to Bayes rule using an exponentiated loss function, shown below:

$$\begin{aligned} \Pr_t(\mu_i \mid a_i) &\propto \Pr(a_i \mid \mu_i) \Pr_{t-1}(\mu_i) \\ &\propto e^{-L(\mu_i)} \Pr_{t-1}(\mu_i) \end{aligned}$$

where L is a binary loss function with a value of 1 if the model incorrectly predicts the observed action and 0 otherwise. We do not test this agent over successive trials, as it has no means of learning a teammate's policy over time. As the model only evaluates the likelihood of the teammate's current target, the belief distribution is reset at the start of each trial.

Types of Teammates

In order to ensure a degree of uncertainty in the paired teammate's behavior, we test the coordinating ad hoc agents with a teammates sampled randomly from a set varying in degrees of consistency of behavior, as follows:

1. Deterministic - This teammate consistently selects its target across trials and pursues it in an identical, deterministic manner every round.

2. Random Target - Here, the teammate begins each trial by uniformly sampling which target it will pursue.

3. Inconsistent - During 90% of the turns, this teammate will pursue its current target while it will select a random action with 10% probability. Furthermore, while the initial target is sampled uniformly, with each step, the teammate may switch targets to evader, $e_j \in E$, with probability proportional to its distance, D, given by $\Pr(e_i \to e_j) \propto 0.2\, D(e_j)$.

The agent does not know the true behavior of the teammate with which it is cooperating; rather, it must learn, infer, or query the teammate's intended actions in order to develop a successful coordinating policy.

9.1.5 Information Over Repeated Trials

Communicating policy information is proposed to handle cases when an agent is uncertain which action a teammate will take. There are two main sources of this uncertainty for the *No Priors* agent:

1. Inconsistency in behavior - Across many observations of a state, a teammate has taken multiple actions.

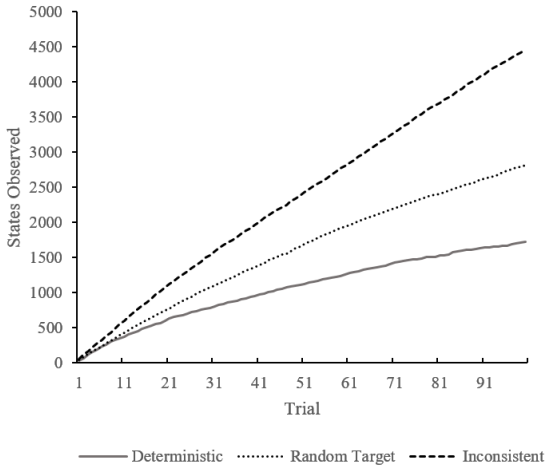

Figure 9.2: Number of unique states observed over successive trials.

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 previously observed policy the teammate is currently following. As an example, consider the teammate beginning in the lower right corner of the maze. Across multiple trials, the collaborating agent observes its teammate either proceed north to pursue the evader in the top right corner or proceed west in pursuit of the bottom left evader. 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. Figure 9.2 presents the average number of unique

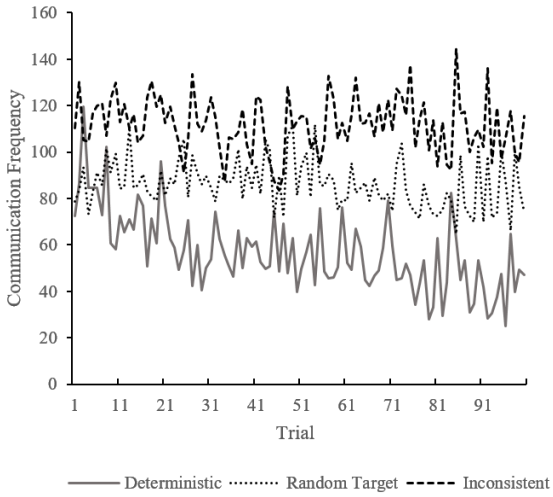

Figure 9.3: Number of queries by the *No Priors* agent over successive trials.

Teammate	Spearman's ρ	p
Deterministic	-0.627	<0.001
Random Target	-0.295	0.003
Inconsistent	-0.130	0.196

Table 9.1: Monotonicity of communicative frequency.

states visited across twenty runs of one hundred successive trials, wherein the *No Priors* agent retains its observations between trials. At each step, the agent selects all queries with positive utility, leaving uncertain any states where communication has no immediate value, given the finite horizon plan. The propensity for the less consistent agents to transition the scenario into new areas of the state space leaves the coordinating agent uncertain, as reflected in the increased frequency of communicative acts, as shown in Figure 9.3.

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 [98], where synced team expectations regarding the status of a task and the individual responsibilities of team members results in lessened conflict and infre-

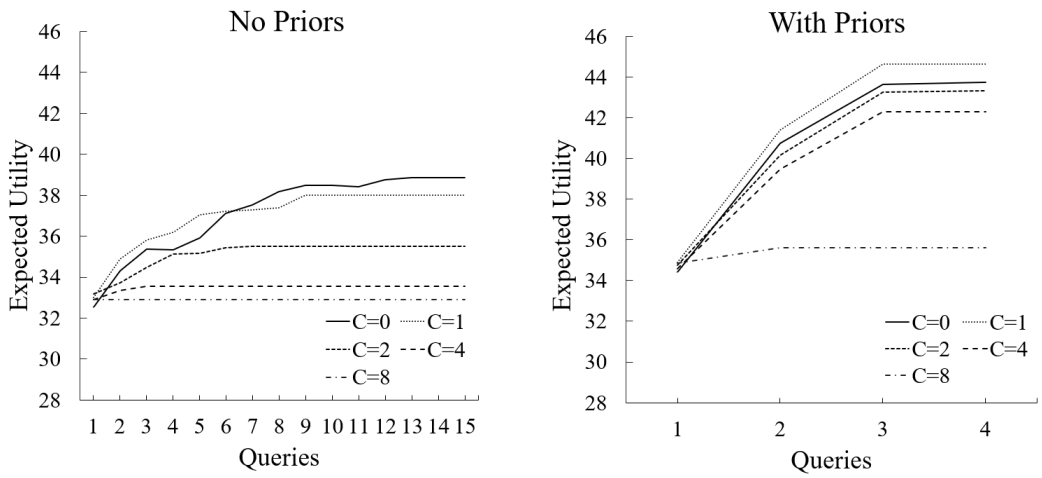

Figure 9.4: Progression of agent's expected utility during Round 1 over successive queries, under various communication costs

quent communication. Table 9.1 contains Spearman's ρ coefficients for the trends displayed in Figure 9.3, indicating the monotonicity of the data. Within the one hundred successive trials, the overall average trends for communicative frequency have significant negative relationships for the Deterministic and Random Target teammate types. However, one hundred trials under the varied performance of the inconsistent teammate appears insufficient for a statistically significant trend.

9.1.6 Cost-restricted Communication

We model the impact of a fixed cost for all communicative acts, though other schemes of assigning costs are possible. Figure 9.4 illustrates the effect of cost on the communication process over successive queries during the first round of coordination. 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. In a contrast of the two modeling approaches, the *With Priors* agent communicates more infrequently, as queries adjust the likelihood of entire

teammate policies, reducing uncertainty for future predictions across all states.

As expected, we observe that increased costs diminish the utilization of communication. In high cost scenarios, agents may only communicate rarely, relying instead on planning under uncertainty with respect to a teammate's behavior. With lower communicative costs, agents exchange information more readily, allowing for reduced uncertainty and increased expected utility.

Table 9.2 displays the success rates for capturing a prey within the time limit as well as statistics for the average reward, as evaluated under varying costs of communication. We observe that the greedy approaches tested in this Chapter were unable to significantly improve the success rates of capture over baseline non-communicating agents. Furthermore, we note significantly worse performance with respect to the cumulative reward, suggesting that agents overutilized communication, likely from overestimating the value of information.

					Reward		
Heuristic	Cost	Trials	Successes	$p_{success}$	Avg.	Std.	p_{util}
With Priors	—	100	89	0.589	45.31	34.01	1.000
With Priors	0	100	80	0.975	27.38	27.11	0.000
With Priors	1	100	87	0.743	4.78	35.88	0.000
With Priors	2	100	85	0.853	-2.09	38.44	0.000
With Priors	4	100	83	0.924	-3.82	39.40	0.000
With Priors	8	100	87	0.743	27.43	40.58	0.001
With Priors	16	100	90	0.500	48.03	34.12	0.573
No Priors	—	100	56	0.557	30.53	35.00	1.000
No Priors	0	100	44	0.967	19.09	30.14	0.014
No Priors	1	100	42	0.983	-38.98	52.08	0.000
No Priors	2	100	54	0.665	-52.89	42.92	0.000
No Priors	4	100	61	0.283	-12.30	38.73	0.000
No Priors	8	100	49	0.871	22.04	32.69	0.078
No Priors	16	100	61	0.283	28.55	34.89	0.689

Table 9.2: Performance of tested agents under varying costs of communication.

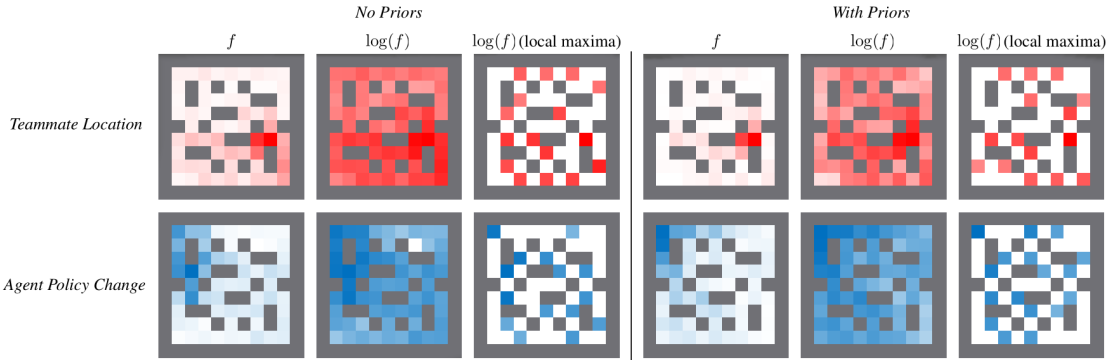

Figure 9.5: Heatmaps for the queries chosen by the agent when coordinating with an unknown teammate. The first row represents the frequencies, f, of the potential teammate locations in the states queried, while the second row depicts where the agent's policy is changed as a result of the queries. Due 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, notably occurring at branchpoints.

9.1.7 Queried States and Policy Changes

When evaluating a potential state query, three elements factor into the value of the communicative act. First, large variance in utility at stake corresponds to increased risk of an uncertain prediction, incentivizing communication. In the tested domain, this primarily occurs at the cusp of a capture. Both teammates must enter the evader's cell to capture it. If the teammate switches targets or performs a random action, it may miss the window for capture, allowing the evader 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 $0 \leq \Pr(\bar{a} \mid s) \leq 1$ as well as transition probabilities (in stochastic domains) $0 \leq T(s, \bar{a}, s') \leq 1$, the value of a

query is biased toward states temporally nearer to the current state. 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 9.5 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 9.5) 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.

Query states for the *With Priors* agent are also presented in Figure 9.5. We observe two apparent points of comparison. First, the *With Priors* agent communicates less frequently, as discussed in the previous section. However, the local maxima for queried states appear in locations in common with those of the *No Priors* agent. This reflects the effect of domain structure independent of the underlying teammate modeling approach used.

9.2 Summary

In this chapter, we proposed several approximation methodologies inspired by the need to reduce the computational requirements of choosing a communication policy. In many respects, the results presented in this initial evaluation demonstrated characteristics of communication policies we would expect, such as the monotonicity of communication with gained information, the impact of cost on the frequency of queries, and the ability to identify decision points which differentiate teammate strategies. Nonetheless, the proposed method failed to improve the performance of the team, perhaps due to the overestimation of the value of information.

The immediate extension to this work, which we introduce in the next two chapters, considers the communication of multiple state-action pairs in succession. Given that it is possible for two states to have no utility for communication individually but have non-zero utility when considered together, many domains may require non-myopic communication policies. This opens up a combinatorial space of potential policy 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. [116], in which the authors successfully employed heuristic techniques to construct sets of observations to share between teammates. Such a technique will not transfer directly to our problem, as we must condition future queries on the teammate responses; however, we similarly explore heuristics in constraining the space of query policies.

Chapter 10

Heuristic Query Evaluators for Search in Information Space

While the full formulation of the communicative process allows for an exhaustive search within the information state space, such an approach is infeasible for domains with moderately large state spaces. The approach of the previous chapter sought to establish a greedy, one-step lookahead as the basis of selecting queries, and it evaluated a query for each state represented in the agent's finite-horizon policy. As we are motivated to consider non-greedy approaches, such as local search, it is natural to consider pruning the space of queries to evaluate, making a trade-off between the breadth and depth of such a search. Under certain conditions, it is possible to provably eliminate queries from consideration, such as when an agent's beliefs are certain on a particular teammate's policy decision. However, such conditions are often rare, providing an insignificant reduction in the space of queries to evaluate. This motivates additional mechanisms by which we can narrow the search space. As such, we outline criteria of desirable heuristic query evaluators for use in the communicative search process and propose a set of candidate heuristics to be evaluated in Chapter 11.

In an effort to avoid confusion over terminology, it is important to stress the distinction between *heuristic query evaluators*—the topic of this chapter—and the *information state heuristic evaluation*, as discussed in Section 8.2, where we defined an admissible heuristic for evaluating the future potential value of information of a given information state. Heuristic query evaluators reduce the branching factor of the search through the information state space, acting as a pruning mechanism in general search problems. The purpose of such evaluators is to remove from consideration many areas of the state space, as deemed irrelevant by the heuristic. In practice, this allows for the search process to probe to deeper horizons of information given a fixed amount of computation, which in our scenario may be necessary for complex coordination efforts that require many policy queries. This is in contrast to heuristic state evaluators, which provide an informed estimate of the future potential utility to be gained from a given state onward. State evaluators provide estimations of the potential *future* value of states yet explored, while query evaluators remove from consideration less useful areas for exploration currently within reach of the search process. The two techniques are commonly used together and have had many powerful results, particularly in the case of AlphaGo Zero [128], which employed a neural network for both state evaluation and policy evaluation, the latter of which functioned as a pseudo-action pruning mechanism by assigning probabilities to actions to bias the search away from actions with small likelihoods of being optimal.

10.1 Desired Characteristics

As the goal of employing heuristics is to eliminate much of the computational burden required to find a good communication policy, it is evident that heuristic query evaluators must both identify a substantially smaller subset of the state

space for analysis and make such a reduction without requiring untenable computational needs in order to identify the subset. As a useful example, a heuristic that requires recomputing the agent's entire policy for each possible query it evaluates may require computation, in the worst case, on the order of $|S||A|$ for every information state evaluated. Given that we're evaluating up to $|S|$ queries and potentially $|A|$ responses at each information state, it is desirable to find a heuristic that selects candidate queries with computation on the order of $|S||A|$ or, in other words, a single pass of the states and actions currently covered in the agent's policy.

10.2 Considerations for Designing Heuristics

A useful strategy for pruning large action spaces is to provide an ordering under some evaluative function, then select only the top k actions. For the purpose assigning approximate ranks to possible policy queries, it is necessary to understand the information available as well as how the information relates to task of computing a successful coordination policy. Consider a scenario in which an agent, currently with beliefs b_0 at a state s_0 evaluates a potential policy query of a state s_t given that it will reach s_t at a future time when it have made one or more observations and arrived at a new a new belief state, b_t. Many useful questions arise from just this subset of information, even if it omits consideration for the remaining reachable world and belief states. How uncertain will the agent be at b_t regarding predictions for s_t? How much utility is at risk of a misprediction? How does knowing $\mu_i(s_t)$ impact the agent's uncertainty over teammate policies with respect to the decision at b_t? Similarly, how is risk or the expected utility at b_t affected? Moreover, many of these lines of inquiry apply to the agent's current decision at b_0. For example, knowing $\mu_i(s_t)$ can impact the likelihoods $\mu_i(s_0)$,

which may in turn induce a policy change for the agent at b_0. In short, the agent considers how the potential for new information changes its policy both in the future *and* at the present. Consequently, to discuss the characteristics and applications of various heuristic query evaluators, we establish broad categorizations of such heuristics.

Information-Theoretic // Decision-Theoretic

The core problem of communicative ad hoc teamwork is two-fold: How does an ad hoc agent make decisions under the uncertainty of teammate policies? How does an agent acquire and utilize information to reduce uncertainty in order to improve its decision-making? From the latter perspective, we look to metrics from *information theory* [127] relating to uncertainty, such as entropy, joint-entropy, and mutual information. Many such metrics have been used in active learning pursuits to select highly-informative data points for labeling [125], much in the same way we desire to use them. Of course, in contrast to much of the classification-based research in active learning, accurate predictions of teammate policies is a secondary goal to that of maximizing the agent's expected utility. It may be the case that certain mispredictions of a teammate's policy in corresponding states may be acceptable, risking relatively little utility, while other states carry a higher variance in the expected payoffs. Consider two agents trapping an evader. While far from the evaders, little utility may be at risk for any single decision along the path(s) toward an evader. However, once in close proximity to the evader, the agents must more closely coordinate their movements to avoid the evader slipping past one of the agents [1], costing the agents the near-term reward of capturing the evader. With this observation, it follows that risk and other evaluations of utility at stake at a given state may be indicators of useful policy queries. We categorize

[1]Recall that it takes both pursuers in the cell with the evader to capture it.

heuristics concerned solely with the uncertainty of predicting teammate behavior as *information-theoretic* while measures incorporating notions of the utility at stake as *decision-theoretic*.

Immediate // Future

Ultimately, an agent querying its teammate for policy information is primarily concerned with its most immediate decision (at b_0/s_0), particularly when it is free to resolve future uncertainties with later communication. It is natural, then, for an agent seek information impacting its current decision, potentially allowing for an opportunity for the agent to change its intended action or confirming its decision by reducing the risk of miscoordination. This implies a relationship between the information gained by querying s_t and the factors of the decision being made at s_0. As such, we refer to metrics proposed on this relationship as *immediate* metrics.

Conversely, as we consider potential query states under the context of a future belief states b_t, given a series of policy observations from s_0 through s_{t-1}, we can construct metrics that take into account factors such as uncertainty and risk *despite* possessing more information at time t. Often, this indicates that such intermediate observations have not been sufficiently informative to make a confident policy decision at b_t. While it may be feasible to clarify a teammate's intentions regarding s_t at a later opportunity, doing so in advance can influence the agent's policy at many intermediate states $(\mathsf{b}_0, \ldots, \mathsf{b}_{t-1})$, affecting the likelihood of visiting s_t. Of course, it is undesirable to recalculate the agent's policy in order to compute such effects, but we can provide estimates of such impacts. Regardless, we refer to metrics relating to information surrounding the decision to be made at s_t under b_t as *future* metrics, though as the majority of the proposed heuristics fall into this category, we will omit this label for conciseness.

10.2.1 State Likelihood Weighting

As Barrett et al. [12] 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, as much of the state space can be considered *unreachable* under the policies of the agent coordinating, effectively pruning off much of the state space. Aside from this consideration, domains may features dynamics of the world wherein actions of the agents have nondeterministic outcomes, e.g. in many robot navigation tasks, it is assumed that with some small probability a robot moving in one direction may actually move in another. In situations where such rare events have low immediate risk, the decisions of the coordinating actions beyond such a transition may not contribute substantially to the agent's policy *prior* to the transition occurring. It is evident that some degree of likelihood may play a role in query selection, as extremely unlikely or even unreachable states may have little importance on the agent's policy decisions relative to more likely outcomes. In order to account for the likelihood of entering a state, we supplement the set of the proposed heuristics with variants for the non-immediate heuristics, incorporating a *state likelihood weighting* component, $w(\mathsf{b}_t, \mathsf{b}_0) = \Pr(\mathsf{b}_t \mid \mathsf{b}_0)$, where we can recursively compute $\Pr(\mathsf{b}_t)$ by

$$\Pr(\mathsf{b}_t) = \sum_{\mathsf{b}_{t-1} \in \mathsf{S}_\mathsf{b}} \sum_{a_0 \in A_0} \mathsf{T}(\mathsf{b}_{t-1}, a_0, \mathsf{b}_t) \Pr(a_0 \mid \mathsf{b}_{t-1}) \Pr(\mathsf{b}_{t-1}) \tag{10.1}$$

Unlike other analysis of state trajectory likelihoods in MDPs, our usage has a unique consideration that plays into our estimation of a state likelihood. Specifically, the agent's policy is not fixed at the stage of communication. In fact, the

purpose of communication is to potentially change the agent's policy. Resolving uncertainty outside of the trajectories resulting from the agent's current policy can incentivize the agent to alter its policy toward trajectories in which the risk has been reduced. It is desirable, then, to adopt a probabilistic policy for the purpose of estimating state likelihoods, as to avoid pruning out state trajectories prematurely. For this, we assign action likelihoods to the coordinating agent's policy as

$$\Pr(a_0 \mid \mathsf{b}_{t-1}) \propto e^{Q(\mathsf{b}_{t-1}, a_0)}$$

$$\text{where } Q(\mathsf{b}_{t-1}, a_0) = \sum_{\mathsf{b}_t \in \mathsf{S}_\mathsf{b}} \mathsf{T}(\mathsf{b}_t, a_0, \mathsf{b}'_t) \left[\mathsf{R}(\mathsf{b}_t, a_0, \mathsf{b}'_t) + \arg\max_{a'_0 \in A_0} \gamma Q(\mathsf{b}'_t, a'_0) \right].$$

10.3 Candidate Heuristics

With the heuristic terminology established, we outline the set of heuristic query evaluators we empirically evaluate in Chapter 11. In order to compare the effectiveness of the heuristics, we establish three initial baseline heuristics.

State Likelihood Each query is assigned its likelihood, according to Equation 10.1.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \Pr(\mathsf{b}_t \mid \mathsf{b}_0)$$

Random Evaluation Each query is assigned a random value, $0 \leq x \leq 1$.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = x, \quad x \sim [0, 1]$$

Weighted Random Each query is assigned a random value, but is further

weighted by its likelihood.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \Pr\left(\mathsf{b}_t \mid \mathsf{b}_0\right) \cdot x, \quad x \sim [0, 1]$$

10.3.1 Information-theoretic Heuristics

From the field of information theory, we employ the concept of *information entropy*, formulated as $E(Y) = -\sum_{y \in Y} \Pr(y) \log \Pr(y)$ for some potential set of events, Y. Information entropy is a measure of the quantity of information contained in a probability distribution, where the uncertainty of an outcome is directly related to the informativeness of witnessing the outcome. Conversely, the more an outcome is certain, the less informative it is to be observed. Within the context of ad hoc teamwork, consider an agent whose belief distribution covers some large number, n, of potential policies for a teammate. If the policies are evenly divided over which of two actions will be taken in a particular state, querying the teammate's policy for the state will rule out one half of the potential policies. Likewise, if all but one potential policy agree on an action, the informativeness of querying the policy for this state is comparatively low, as we would expect in $\frac{n-1}{n}$ cases that the majority consensus action would be the responses, which in turn rules out only a single policy. Even though in the $\frac{1}{n}$ instances the rare response occurs, the agent gains much information, the likelihood of this result is small enough such that the total expected information is low.

For the purpose of designing heuristics, we observe two sets of information from which we can reason about potential information gain. First, the agent maintains a distribution over the policies a teammate. Secondly, from the distribution of policies, we construct distributions of action likelihood given a state.

Action Information Entropy The information entropy for the uncertainty regarding a teammate's actions in a state, s_t, given the beliefs, b_t, it holds when it reaches the state.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = E\left(a_i \mid \mathsf{b}_t\right) = -\sum_{a_i \in A_i} \Pr(a_i \mid \mathsf{b}_t) \log \Pr(a_i \mid \mathsf{b}_t)$$

Weighted Action Information Entropy As above, but weighted with the state likelihood. b_t, it may hold when it reaches the state.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \Pr\left(\mathsf{b}_t \mid \mathsf{b}_0\right) E\left(a_i \mid \mathsf{b}_t\right)$$

Δ Policy Entropy The expected change in entropy over potential policies if the teammate policy at s_t with belief b_t were to be queried.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = E\left(\mu_i \mid \mathsf{b}_t\right) - \mathbb{E}\left[E\left(\mu_i \mid \mathsf{b}'_t\right) \mid \mu_i(s_t)\right]$$

Weighted Δ Policy Entropy As above, but weighted by the state likelihood.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \Pr\left(\mathsf{b}_t \mid \mathsf{b}_0\right) \left[E\left(\mu_i \mid \mathsf{b}_t\right) - \mathbb{E}\left[E\left(\mu_i \mid \mathsf{b}'_t\right) \mid \mu_i(s_t)\right]\right]$$

Immediate Δ Policy Entropy As opposed to the previous heuristics, we consider how resolving uncertainty at s_t impacts the uncertainty at s_0.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = E\left(\mu_i \mid \mathsf{b}_0\right) - \mathbb{E}\left[E\left(\mu_i \mid \mathsf{b}'_0\right) \mid \mu_i(s_t)\right]$$

10.3.2 Decision-theoretic Metrics

As we discussed earlier, under certain conditions, uncertainty is acceptable, and utility should be taken into account in order to evaluate the risk of mispre-

dictions. Toward this end, we adopt two perspectives: how much utility is at risk due mispredicting teammate policies—the potential error due to uncertainty—and how much utility is to be gained by knowing the truth, i.e. the value of information.

Mean Absolute Error The expected error of a prediction made at s_t given b_t.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \mathbb{E}_{\mu(s_t)}\Big[\ \Big|V(\mathsf{b}'_t) - V(\mathsf{b}_t)\Big|\ \Big]$$

Weighted Mean Absolute Error As above, but weighted by the state likelihood.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \Pr\left(\mathsf{b}_t \mid \mathsf{b}_0\right) \mathbb{E}_{\mu(s_t)}\Big[\ \Big|V(\mathsf{b}'_t) - V(\mathsf{b}_t)\Big|\ \Big]$$

Mean Squared Error The expected squared error of a prediction made at s_t given b_t.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \mathbb{E}_{\mu(s_t)}\left[\ \left(V(\mathsf{b}'_t) - V(\mathsf{b}_t)\right)^2\ \right]$$

Weighted Mean Squared Error As above, but weighted by the state likelihood.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \Pr\left(\mathsf{b}_t \mid \mathsf{b}_0\right) \mathbb{E}_{\mu(s_t)}\left[\ \left(V(\mathsf{b}'_t) - V(\mathsf{b}_t)\right)^2\ \right]$$

While the true value of information requires recomputing the agent's beliefs and policy for all future states, we can create an approximate to this value by only examining how the beliefs and policy would change at the decision around $\mathsf{b}t$, holding all other beliefs and policy choices constant. With this approximation, we construct the following heuristic evaluation functions:

Approximate Value of Information The approximate value of information for knowing a teammate's policy at s_t.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \mathbb{E}\left[\max_{a_0 \in A} Q(\mathsf{b}_t, a_0) \mid \mu(s_t)\right] - V(\mathsf{b}_t)$$

Weighted Approximate Value of Information As above, but weighted by the state likelihood.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \Pr\left(\mathsf{b}_t \mid \mathsf{b}_0\right) \mathbb{E}\left[\max_{a_0 \in A} Q(\mathsf{b}_t, a_0) \mid \mu(s_t)\right] - V(\mathsf{b}_t)$$

Immediate Approximate Value of Information The approximate value of information the agent at the current decision at s_0.

$$\mathrm{H}(\mathsf{b}_t, \mathsf{b}_0) = \mathbb{E}\left[\max_{a_0 \in A} Q(\mathsf{b}_0, a_0) \mid \mu(s_t)\right] - V(\mathsf{b}_0)$$

10.4 Summary

In this chapter, we have described a few motivating criteria for evaluating the quality of a query. While we have proposed a number of heuristics designed with these criteria in mind, there is much potential for further investigation in this direction. The heuristics proposed here have been based only on information present within a generalized model of the Policy Communication Decision Problem and were intended for applications to general domains. For practical applications, domain-specific heuristics can potentially improve the ability to select crucial queries for coordination. As an example, we see in the query results from the two agent pursuit problem that branches in a maze are common queries.

Clearly, this is a result of patterns in optimal policies where for states with the teammate is in a corridor have similar behavior, i.e. the teammate continues along the corridor at each turn. In contrast, forks of the maze represent states where groups of policies have divergent decisions. It is possible to design heuristics that identify these obvious decision points without needing to evaluate much of the state space. An advantage an agent with much experience operating with ad hoc teams in a domain is that it learns the distribution of team policies in practice, then identifies common policy divergences from this experience.

Furthermore, it may be possible to create a learned heuristic. One could employ reinforcement learning to attempt to converge to the correct the value of a query in an given information state over time. Furthermore, one could attempt to classify states as good candidates for policy queries directly from the state description. In domains are represented by a vector of values, such as the vector of pixel values used for many of the Atari domains [17], there is a potential to identify patterns in the data which are features correlating with high informativeness.

With an initial set of heuristics, we move to the empirical analysis of the Policy Communication Decision Problem in the multiagent pursuit domain. We will show how heuristics, together with a local search approach, can provide improved coordination while retaining a tractable degree of computation.

Chapter 11

Empirical Evaluation of Heuristic Query Evaluators

In the previous chapter, we proposed a set of heuristic query evaluators with the purpose of pruning the space of queries state to a manageable set, the particular cardinality of which is left up to the application or domain-specific optimization. For our purposes, we will demonstrate the effects of adjusting various algorithm and domain hyperparameters, with the goal of showing how the choice of each impacts an agent's ability to identify policy information worth querying. We keep such algorithmic considerations as domain-independent as possible, though application to other domains may necessitate further tuning. We discuss general trends and lessons learned both here and in Chapter 12. With this in mind, we address the following research questions with this evaluation:

- *How do each of the heuristic query evaluators perform under various parameterizations of the search branching factor and overall search budget?* (Section 11.2.1)
- *How do each of the heuristics perform with varying levels of agent experi-*

ence? (Section 11.2.2)

- *How does the variety of teammate policies affect the utilization of communication, due to the difficulty of learning priors for the teammate population?* (Section 11.2.3)

- *How does the cost of communication impact the use and effectiveness of communication?* (Section 11.2.4)

- *How does the structure of a domain create need for explicit information-gathering?* (Section 11.2.5)

11.1 Experimental Setup

As before, we will use the two-agent pursuit domain as our target application domain. For this evaluation, we use the maze in Figure 11.1, with a team reward of 100 if an evader is caught and 0 otherwise. We cap the maximum number of turns to 7. In contrast to the experimental setups of in Chapters 5 and 9, this constraint puts pressure on the team to coordinate precisely, with low tolerance for mistakes. Moreover, the shortened time constraint limits the reachable state space, allowing us to sample from the complete space of optimal teammate policies, of which there are 361 individual policies which achieve the joint goal, and also to measure the effect of variance among sampled teammate policies, which will discuss in Section 11.2.3.

Furthermore, for reach trial, the agent attempts collaboration with a newly sampled teammate. While repeated trials with the same teammate are useful for demonstrating the capability of models to be learned from observation, initial trials typically exhibit the highest rate of failure, due to small number of observations

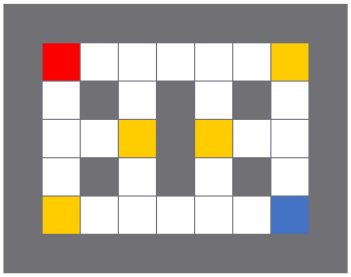

Figure 11.1: Maze used for communication search evaluation. The coordinating agent is represented in blue, with the teammate in red.

a coordinating agent possesses. As such, initial attempts at coordination with a new teammate have the highest potential for improvement due to communication.

11.1.1 Sampling Teammate Policies

As we are working with a scenario in which we can fully compute the space of teammate policies, we can sample new policies according to any distribution we impose over the policy space. As a default, we will sample policies uniformly from the set of policies which achieve the optimal value function, i.e. $V_{\mu}(s) = V^{*}(s)\ s \in S$. In Section 11.2.3, we will vary the space of policies from which teammate strategies are sampled, showing the effect of coordinating with both large and small varieties of teammates.

11.1.2 Beliefs Over Teammate Policies

With a teammate policy generation mechanism in place, we can simulate the experience of an agent by sampling from this set. To simulate an agent's past experience coordinating with teammates, we sample one teammate policy per *episode* of experience. We quantify an agent's total past experience by the total episodes of past experience. Notably, as we sample policies with replacement, policies may

be sampled more than once, forming a distribution over the relative frequencies of the policies seen. As one of the primary initiatives of ad hoc teamwork centers around coordinating with *novel* teammate strategies, we are interested in scenarios where the agent's experience has not covered the entire space of policies or has not sampled sufficiently many times such that it has perfectly learned the relative frequencies of each policy. For this reason, we will primarily use sets of past experience smaller than the set of all possible optimal teammate policies (here, 361). As a result, the initial belief distribution of the agent is incomplete, not possessing any experience coordinating with certain teammate policies.

Rather than assuming a uniform prior over teammate policies or constructing a prior by other measures as in [8], we employ priors generated from a Chinese Restaurant Process [103]. Through sampling, a Chinese Restaurant Process (from here, CRP) is a discrete-time stochastic process capable of approximating a probability distribution over discrete items in a potentially infinite set. Given a count, $c(\mu_i)$, for each policy observed previously and a concentration parameter[1], $\alpha \geq 0$, we estimate the probability of a policy by

$$\Pr(\mu_i) = \frac{c(\mu_i)}{\alpha + \sum_j c(\mu_j)}$$

and the probability of encountering a novel policy by

$$\Pr(\mu_?) = \frac{\alpha}{\alpha + \sum_j c(\mu_j)} \,.$$

With sufficiently many samples, this estimation approximates the true underlying distribution of policies. By reserving probability for the space of yet encountered policies, the distribution eschews the need of alternative belief revi-

[1]We discuss the computation of α in Appendix A.

sion strategies. Rather, the posterior assigns a probability of 0 is assigned to any policy that does not agree with the observation of the teammate policy and 1 otherwise. In situations where an agent has ruled out every known policy, $\Pr(\mu_?) = 1$, concluding the currently observed policy is from the set of unknown policies. For this possibility, we predict actions uniformly from the set of all teammate actions, even if the action is not part of any optimal joint policy for the team. This permits the consideration that a teammate performs sub-optimal actions. In practice, the default uniform prediction can be replaced with any distribution, such that every action has a non-zero likelihood. With a default prediction scheme in place, given a belief distribution over policies, we compute the cumulative action likelihoods as

$$\Pr(a_i \mid s) = |A_i|^{-1} \Pr(\mu_?) + \sum_{\substack{\mu_i \in A_i^S \\ \mu_i(s) = a_i}} \Pr(\mu_i).$$

Under this construction, every teammate policy and every teammate action prediction has a non-zero probability. However, with many observations consistent with a known policy, theses likelihoods can become arbitrarily small as the agent becomes more confident in the predictions of a policy.

11.1.3 Computing a Policy

Using the POMDP framework outlined in Chapter 2, the agent computes an individual policy through solving the belief MDP using a prior over policies from the CRP outlined previously. To solve the MDP, we use an implementation of Trial-based Heuristic Tree Search (THTS) [74], a heuristic, sample-based search algorithm which generalizes a family of similar search algorithms, such as Monte Carlo Tree Search [34] and LAO* [60]. As with MCTS, the algorithm uses Upper

Confidence Bounds for Trees (UCT) [75] to guide the search. As this search approach is sample-based, we will refer to the number of samples as the *iterations* of the algorithm executed. The state value estimates and resulting policy asymptotically converge to the optimal values and policy, respectively, with the increase of search iterations. THTS possesses two primary advantages over the MCTS implementations from Chapters 5 and 9. First, it utilizes dynamic programming due to the Markov assumption of the MDPs; this changes the structure of the search from a tree search to a graph search, wherein multiple trajectories may converge to an identical state. In domains where convergence occurs, the search process more efficiently covers the state space by avoiding duplicate computation. Secondly, the algorithm computes value estimates using full Bellman backups (see Equation 2.1) rather than simple sample-based averages, yielding more accurate policy values.

Once a coordination policy has been computed, the policy is fed into a separate instance of the same solver in order to compute a communication policy, using the MDP construction of the problem from Chapter 7. While both policies are computed with the same solver, the parameterization of each search process differs, which we will outline within the set of all hyperparameters of the evaluation.

11.1.4 Overview of Hyperparameters

For the body of experiments in this chapter, there are several test hyperparameters, a portion of which will remain fixed, as they do not address the research questions outlined earlier. The fixed parameters are assigned as follows:

- Planning Iterations (Coordination Policy) - The number of iterations per planning step. [default=500]
- Coordination policy search heuristic - The state evaluation heuristic for the

coordination policy, computed as a function of agent location, a, teammate location, s, evader locations, $r \in R$, the current turn, t, and the Manhattan distance function, M.

$$h_S(s) = \begin{cases} 100 & \exists r \in R,\ M(a,r) < 7 - t \text{ and } M(s,r) < 7 - t \\ 0 & \text{otherwise} \end{cases}$$

- Communication Policy Search Heuristic - The state evaluation heuristic for the communication policy. Here, $h_{\mathcal{I}}(I) = 0$.

Additionally, we adjust and measure the effects of the following experimental variables:

- Heuristic Query Evaluators - The set of query evaluators proposed in Chapter 10. All experiments evaluate each heuristic.
- Communication Branching Factor - After heuristically evaluating each potential query, the search process will select the top k queries. [default=5]
- Planning Iterations (Communication) - The number of iterations per planning step while computing the communication policy. [default=10]
- Experience - The total number of policies the agent has observed from past experience. [default=100]
- Maximum Unique Teammate Policies - An upper bound on the number of unique teammate policies the agent may encounter. [default=∞]
- Cost of Communication - A flat cost assigned to each query. [default=5]

11.2 Empirical Results

11.2.1 Communication Search Parameters

Due to the exponentially large search space for the communication decision problem, we limit the branching factor of the search by selecting only the top k queries at each information state in the communication policy. Under the constraint of a search budget—here the number of iterations of the sample-based search—this pruning allows the search to probe deeper into larger information sets than in a non-pruned search. Furthermore, as the evaluation of each information state requires the recomputation of the agent's beliefs and policy, it is beneficial to minimize the search iterations. In effect, the communication partial policies that result are substantially smaller than the search policies recomputed. Complete results for each heuristic query evaluator with search parameters $k \in \{1, 3, 5\}$ and $Iter \in \{1, 10, 20\}$ are presented in Appendix B.

We compare the performance of the various parameterizations by two metrics: success rate for coordinating and the average reward earned by the team. It is useful to distinguish between these metrics, as a team may be highly successful but overutilize communication, resulting in a lower overall reward. The results report p values for both metrics across each heuristic, as compared to a baseline agent with identical experience and coordination planning iterations but without the ability to communicate. We note that the random query evaluator did not result in significant improvement in either metric compared to the non-communicating baseline.

A selection of the full results are presented here. Table 11.1 provides results from the most constrained search parameterization[2], i.e. $Iter = 1,\ k = 1$. Addi-

[2]We discuss the comparison of these pruned, greedy results with respect to the results of Chapter 9 in Section 11.3.2.

tionally, Table 11.2 contains results for the default values of the branching factor ($k = 5$) and search iterations ($Iter = 10$), as used in the remaining experiments.

From the results, we observe several apparent trends:

- The local, entropy-based heuristics—local action information entropy and local policy information entropy—never significantly outperform the non-communicating baseline.

- The local, error-based heuristics—local absolute error and local mean squared error—have somewhat mixed results, able to achieve significant improvements on success rate but often unable to improve on the team's average reward. This suggests the communication policy is able to obtain enough information to coordinate but queries its teammate to such an extent that the cost from queries offset the gains in success.

- The local value of information improved on the baseline performance in every configuration.

- With the exception of the weighted random heuristic, the set of weighted heuristics achieved significant improvements for both success and average reward

- The weighted random heuristic was unable to communicate adequately with a branching factor, $k = 1$, yet it mirrored the improvements of the weighted heuristics for $k \in \{3, 5\}$. This demonstrates risk of pruning the queries too aggressively, as the randomness of the heuristic often failed to find informative queries as its top pick, yet it performed well with the best 3 explored.

- Both immediate heuristics, immediate policy entropy and immediate value of information, similarly outperformed the baseline, emphasizing the impact

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	25	1.000	50.00	50.51	0.001
Mean Absolute Error	50	36	0.923	68.30	46.47	0.113
Mean Squared Error	50	28	0.999	52.50	51.34	0.002
Δ Policy Entropy	50	30	0.996	60.00	49.49	0.015
Approx. Value of Info.	50	50	**0.001***	96.60	2.36	**0.011***
Weighted Action Entropy	50	50	**0.001***	96.40	2.27	**0.012***
Weighted Mean Abs. Error	50	50	**0.001***	97.40	2.52	**0.007***
Weighted Mean Sq. Error	50	50	**0.001***	93.30	4.59	**0.046***
Weighted Δ Policy Ent.	50	50	**0.001***	97.20	2.51	**0.008***
Weighted Approx. VoI	50	50	**0.001***	97.10	2.49	**0.008***
Immediate Policy Ent.	50	50	**0.001***	96.80	2.42	**0.010***
Immediate Value of Info.	50	50	**0.001***	96.80	2.42	**0.010***
Uniform Random	50	33	0.980	66.00	47.85	0.069
State Likelihood	50	50	**0.001***	96.40	2.27	**0.012***
Weighted Uniform Random	50	41	0.602	80.00	37.93	0.795

Table 11.1: Heuristics evaluation with communication branch factor of 1 and 1 iteration(s) per search step. * denotes significant improvement over the baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	30	0.996	60.00	49.49	0.015
Mean Absolute Error	50	49	**0.008***	92.60	14.92	0.076
Mean Squared Error	50	49	**0.008***	92.40	14.99	0.082
Δ Policy Entropy	50	32	0.988	64.00	48.49	0.043
Approx. Value of Info.	50	50	**0.001***	96.60	2.36	**0.011***
Weighted Action Entropy	50	50	**0.001***	96.70	2.39	**0.010***
Weighted Mean Abs. Error	50	50	**0.001***	97.10	2.49	**0.008***
Weighted Mean Sq. Error	50	50	**0.001***	96.90	2.45	**0.009***
Weighted Δ Policy Ent.	50	50	**0.001***	94.10	4.81	**0.033***
Weighted Approx. VoI	50	50	**0.001***	96.90	2.45	**0.009***
Immediate Policy Ent.	50	50	**0.001***	96.60	2.36	**0.011***
Immediate Value of Info.	50	50	**0.001***	93.70	4.72	**0.039***
Uniform Random	50	39	0.773	77.50	42.33	0.581
State Likelihood	50	50	**0.001***	97.00	2.47	**0.009***
Weighted Uniform Random	100	100	**0.000***	96.10	3.14	**0.013***

Table 11.2: Heuristics evaluation with communication branch factor of 5 and 10 iteration(s) per search step. * denotes significant improvement over baseline.

knowing a teammate's policy at a future state has on the agent's decision at the current state.

11.2.2 Past Experience

In contrast to manually specified model-based approaches, the agents tested here instead learn a distribution of the population of teammate policies. The accuracy of this learned prior is predicated on the past experience of the agent and impacts the ability for an agent to correctly predict inferred actions of its teammates.

Recall that for the tested maze, we sample teammate policies from the set of 361 individual policies that form part of an optimal joint policy. We test the agent under four levels of experience, $exp \in \{0, 10, 100, 1000\}$, and present results for 0 episodes and 10 episodes of past experience in Tables 11.3 and 11.4, respectively. The remaining results are found in Appendix B.

The two experience levels included here illustrate an important result for communicating agents. With relative inexperience, agents that communicate can coordinate more effectively than non-communicating agents. This is not true in the tested case with 0 episodes of experience, however. Yet, we observe that the non-communicating baseline does not improve in success or average reward with the first 10 episodes of coordination. The communication process better leverages the small set of experience, bootstrapping collaboration despite the inaccurate prior.

At higher experience levels (Appendix B), the non-communicating baseline significantly improves its performance over its low experience counterparts. Nevertheless, the communicating agents achieve better success rates and average rewards, despite the baseline possessing increasingly accurate learned priors. This continued advantage demonstrates the capability of communication to improve

Heuristic	Experience	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	0	50	28	—	56.00	50.14	—
Action Entropy	0	100	63	0.257	63.00	48.52	0.417
Mean Absolute Error	0	50	30	0.420	60.00	49.49	0.689
Mean Squared Error	0	50	33	0.206	66.00	47.85	0.310
Δ Policy Entropy	0	50	27	0.656	54.00	50.35	0.843
Approx. Value of Info.	0	50	28	0.580	56.00	50.14	1.000
Weighted Action Entropy	0	50	35	0.107	70.00	46.29	0.150
Weighted Mean Abs. Error	0	50	29	0.500	58.00	49.86	0.842
Weighted Mean Sq. Error	0	50	36	0.072	72.00	45.36	0.097
Weighted Δ Policy Ent.	0	50	37	**0.046***	74.00	44.31	0.060
Weighted Approx. Vol	0	50	30	0.420	60.00	49.49	0.689
Immediate Policy Ent.	0	50	31	0.342	62.00	49.03	0.547
Immediate Value of Info.	0	50	31	0.342	62.00	49.03	0.547
Uniform Random	0	50	31	0.342	62.00	49.03	0.547
State Likelihood	0	50	31	0.342	62.00	49.03	0.547
Weighted Uniform Random	0	100	70	0.065	70.00	46.06	0.101

Table 11.3: Agent coordinating with 0 episodes of past experience. * denotes significant improvement over baseline.

Heuristic	Experience	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	10	50	29	—	58.00	49.86	—
Action Entropy	10	100	67	0.183	66.70	47.71	0.309
Mean Absolute Error	10	50	50	**0.000***	92.20	5.64	**0.000***
Mean Squared Error	10	50	49	**0.000***	90.80	14.33	**0.000***
Δ Policy Entropy	10	50	34	0.204	67.70	47.15	0.320
Approx. Value of Info.	10	50	50	**0.000***	90.60	6.67	**0.000***
Weighted Action Entropy	10	50	49	**0.000***	92.90	13.71	**0.000***
Weighted Mean Abs. Error	10	50	50	**0.000***	93.30	5.31	**0.000***
Weighted Mean Sq. Error	10	50	50	**0.000***	89.30	9.26	**0.000***
Weighted Δ Policy Ent.	10	50	45	**0.000***	84.80	30.44	**0.002***
Weighted Approx. Vol	10	50	50	**0.000***	90.70	6.47	**0.000***
Immediate Policy Ent.	10	50	41	**0.008***	76.00	39.03	**0.047***
Immediate Value of Info.	10	50	50	**0.000***	91.80	5.87	**0.000***
Uniform Random	10	50	40	**0.015***	75.00	43.69	0.073
State Likelihood	10	50	49	**0.000***	86.20	15.24	**0.000***
Weighted Uniform Random	10	100	93	**0.000***	85.45	25.44	**0.001***

Table 11.4: Agent coordinating with 10 episodes of past experience. * denotes significant improvement over baseline.

coordination when the agent has learned the space of teammate policies well.

11.2.3 Population of Teammates

As noted previously, for each trial, a teammate is sampled from the set of 361 individual policies, which directly relates to the uncertainty inherent to the prior an agent learns over the population of teammates. We are interested in the effect of such variance over policies, as agents coordinating with teammates drawn from a smaller set require fewer samples to estimate the distribution of policies in a population. To measure this effect, we institute a cap on the number of unique policies in an underlying teammate population, though this information is not provided to the agent. We test the population caps of 5, 25, and 125 under agent experience levels, 10, 100, and 1000.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	87	—	87.00	33.80	—
Action Entropy	150	133	0.418	88.10	31.80	0.797
Mean Absolute Error	50	50	0.004	94.80	7.28	**0.029***
Mean Squared Error	50	50	0.004	92.00	10.64	0.179
Δ Policy Entropy	50	45	0.404	87.60	30.54	0.913
Approx. Value of Info.	50	50	0.004	91.90	13.09	0.206
Weighted Action Entropy	50	46	0.267	86.90	29.84	0.985
Weighted Mean Abs. Error	50	50	0.004	92.30	10.46	0.153
Weighted Mean Sq. Error	50	50	0.004	91.50	11.35	0.231
Weighted Δ Policy Ent.	100	96	**0.020***	91.90	20.25	0.215
Weighted Approx. VoI	50	50	0.004	94.30	5.98	**0.038***
Immediate Policy Ent.	88	88	**0.000***	83.24	29.07	0.413
Immediate Value of Info.	50	50	0.004	94.50	12.75	0.052
Uniform Random	50	48	0.069	89.50	22.75	0.593
State Likelihood	50	50	0.004	91.70	11.50	0.212
Weighted Uniform Random	100	98	**0.003***	93.10	14.72	0.100

Table 11.5: Agent coordinating with 10 experience with 5 maximum unique teammate policies. * denotes significant improvement over baseline.

The results shown in Table 11.5 indicate that coordinating with teammates

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	73	—	73.00	44.62	—
Action Entropy	150	108	0.623	71.40	45.41	0.783
Mean Absolute Error	50	50	0.000	90.60	7.60	**0.000***
Mean Squared Error	50	50	0.000	91.90	7.42	**0.000***
Δ Policy Entropy	50	35	0.721	68.50	47.62	0.579
Approx. Value of Info.	50	49	0.000	89.50	15.66	**0.001***
Weighted Action Entropy	50	49	0.000	92.40	13.71	**0.000***
Weighted Mean Abs. Error	50	50	0.000	82.10	16.94	0.074
Weighted Mean Sq. Error	50	50	0.000	88.00	14.07	**0.003***
Weighted Δ Policy Ent.	50	46	0.004	87.80	26.27	**0.012***
Weighted Approx. VoI	50	50	0.000	90.90	5.41	**0.000***
Immediate Policy Ent.	50	37	0.530	66.80	44.02	0.420
Immediate Value of Info.	50	50	0.000	90.30	7.59	**0.000***
Uniform Random	50	42	0.096	78.40	37.57	0.438
State Likelihood	50	50	0.000	88.40	6.88	**0.001***
Weighted Uniform Random	100	96	**0.000***	89.55	20.21	**0.001***

Table 11.6: Agent coordinating with 10 experience with 125 maximum unique teammate policies. * denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	74	—	74.00	44.08	—
Action Entropy	150	107	0.727	71.27	45.33	0.635
Mean Absolute Error	50	50	0.000	80.70	11.43	0.156
Mean Squared Error	100	100	**0.000***	82.85	9.78	0.053
Δ Policy Entropy	50	36	0.679	71.90	45.52	0.788
Approx. Value of Info.	50	50	0.000	92.80	3.52	**0.000***
Weighted Action Entropy	100	89	**0.005***	86.15	30.74	**0.025***
Weighted Mean Abs. Error	100	100	**0.000***	83.45	11.07	**0.040***
Weighted Mean Sq. Error	50	50	0.000	82.70	10.98	0.065
Weighted Δ Policy Ent.	50	44	0.036	85.90	32.76	0.065
Weighted Approx. VoI	50	50	0.000	91.80	5.23	**0.000***
Immediate Policy Ent.	100	92	**0.001***	81.05	27.20	0.175
Immediate Value of Info.	50	50	0.000	95.00	0.00	**0.000***
Uniform Random	100	79	0.252	65.70	42.23	0.176
State Likelihood	50	50	0.000	91.40	4.74	**0.000***
Weighted Uniform Random	50	49	0.000	84.50	16.48	**0.037***

Table 11.7: Agent coordinating with 1000 experience with 125 maximum unique teammate policies. * denotes significant improvement over baseline.

sampled from a small population substantially reduces the failure rate and average reward, as compared to the results with more varied teammate populations. As a consequence, we do not observe many instances in which communication provides an advantage over the non-communicating baseline, which on its own achieves a high degree of success. As a point of comparison, consider the results in Tables 11.6 and 11.7, in which communicative agents surpass the baseline across a majority of the heuristics tested. Aggregating the results here and in Appendix B, we observe that under population caps of 5, 25, and 125, communicative agents surpass the baseline in 4, 25, and 29 instances (of 45 total heuristic/experience combinations), respectively. This illustrates the advantage of utilizing communication to reduce increasing uncertainty from the underlying population.

Less apparent is the significance of experience, as under each population cap, correlations of experience and success are weak or nonexistent. However, we observe that, in aggregate, significant improvements from communicative agents occur more frequently with low agent experience. In scenarios in which the agent have sampled 10 episodes of experience, communication proved beneficial in 23 of 45 configurations (15 heuristics, 3 population caps), as opposed to 16 and 14 instances for 100 and 1000 episodes, respectively.

11.2.4 Cost of Communication

As with much existing work in decision-theoretic communication [108, 10], we are interested in the impact of cost on the use and effectiveness of communication. Here, we test four costs, $C \in \{1, 5, 10, 99\}$. Unsurprisingly, under the highest cost, the agents do not elect to communicate. Moreover, mirroring the mixed results in Section 11.2.1, the unweighted error-based heuristics do not achieve significant results once $C \geq 10$, failing to find short sequences of queries whose value

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	37	—	74.00	44.31	—
Action Entropy	100	61	0.962	61.00	49.02	0.105
Mean Absolute Error	50	50	**0.000***	92.00	4.04	**0.006***
Mean Squared Error	50	50	**0.000***	92.00	5.15	**0.006***
Δ Policy Entropy	50	36	0.674	72.00	45.36	0.824
Approx. Value of Info.	50	50	**0.000***	96.40	2.27	**0.001***
Weighted Action Entropy	50	50	**0.000***	93.80	4.58	**0.003***
Weighted Mean Abs. Error	50	50	**0.000***	96.60	2.36	**0.001***
Weighted Mean Sq. Error	50	50	**0.000***	97.00	2.47	**0.001***
Weighted Δ Policy Ent.	50	50	**0.000***	95.00	4.95	**0.002***
Weighted Approx. VoI	50	50	**0.000***	96.30	2.22	**0.001***
Immediate Policy Ent.	50	50	**0.000***	96.50	2.31	**0.001***
Immediate Value of Info.	50	50	**0.000***	94.80	4.84	**0.002***
Uniform Random	50	34	0.811	66.90	46.85	0.438
State Likelihood	50	50	**0.000***	96.80	2.42	**0.001***
Weighted Uniform Random	50	50	**0.000***	95.30	4.09	**0.001***

Table 11.8: Agent coordinating with communication cost $C(q) = 5$. * denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	37	—	74.00	44.31	—
Action Entropy	50	33	0.862	66.00	47.85	0.388
Mean Absolute Error	50	34	0.811	64.80	45.46	0.308
Mean Squared Error	50	30	0.956	59.20	48.98	0.116
Δ Policy Entropy	50	30	0.956	60.00	49.49	0.139
Approx. Value of Info.	50	50	**0.000***	94.00	4.95	**0.003***
Weighted Action Entropy	50	50	**0.000***	93.60	4.85	**0.003***
Weighted Mean Abs. Error	50	50	**0.000***	94.00	4.95	**0.003***
Weighted Mean Sq. Error	100	99	**0.000***	92.90	10.66	**0.004***
Weighted Δ Policy Ent.	50	50	**0.000***	93.80	4.90	**0.003***
Weighted Approx. VoI	50	50	**0.000***	92.80	4.54	**0.004***
Immediate Policy Ent.	100	100	**0.000***	93.20	4.69	**0.004***
Immediate Value of Info.	50	50	**0.000***	94.20	4.99	**0.002***
Uniform Random	50	33	0.862	66.00	47.85	0.388
State Likelihood	50	50	**0.000***	93.40	4.79	**0.003***
Weighted Uniform Random	50	47	**0.006***	89.20	23.37	**0.035***

Table 11.9: Agent coordinating with communication cost $C(q) = 10$. * denotes significant improvement over baseline.

exceeds the moderate cost. The remaining heuristics achieve results mirroring those in Section 11.2.1, with significant improvement by weighted and immediate heuristics.

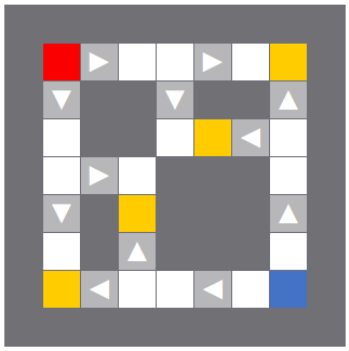

Figure 11.2: Maze structured such that communication is unnecessary. The coordinating agent is represented in blue, with the teammate in red.

11.2.5 Structure of Domain

Finally, while we have thus far shown the effectiveness of communication in the tested domain, it is worthwhile to provide an example where communication is unnecessary. For this purpose, we further tested the heuristics across varying degrees of past experience on the maze presented in Figure 11.2. This maze was specifically designed to force the teammate to make important decisions *prior* to the corresponding decisions of the coordinating agent. In this manner, the agent is always guaranteed to have observed these crucial policy decisions. We further observe that communicating and non-communicating agents alike are able to coordinate with high success, even with no past experience, as shown in Table 11.10. This demonstrates a correspondence between the structure of the domain and the need for communication, which has been a topic largely omitted from existing work.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	48	—	96.00	19.79	—
Action Entropy	150	144	0.638	95.60	19.64	0.902
Mean Absolute Error	50	43	0.985	84.30	35.08	0.043
Mean Squared Error	50	45	0.944	88.60	30.61	0.155
Δ Policy Entropy	100	89	0.967	88.05	31.21	0.060
Approx. Value of Info.	50	48	0.691	94.90	19.86	0.782
Weighted Action Entropy	50	48	0.691	95.00	19.92	0.802
Weighted Mean Abs. Error	100	94	0.812	91.90	23.67	0.266
Weighted Mean Sq. Error	50	44	0.970	85.80	32.31	0.061
Weighted Δ Policy Ent.	50	46	0.898	89.60	27.10	0.181
Weighted Approx. VoI	50	48	0.691	94.20	20.11	0.653
Immediate Policy Ent.	50	47	0.819	93.50	23.93	0.571
Immediate Value of Info.	50	45	0.944	87.20	29.92	0.086
Uniform Random	50	46	0.898	91.40	27.29	0.337
State Likelihood	50	47	0.819	91.10	23.95	0.268
Weighted Uniform Random	100	87	0.985	85.60	33.50	0.019

Table 11.10: Results from tests with the maze depicted in Figure 11.2. Each agent tested was initialized with 0 episodes of past experience. * denotes significant improvement over baseline.

11.3 Summary

In this chapter, we have empirically evaluated the proposed heuristic-based approach to computing communication policies for the communicative ad hoc teamwork problem. From these results, we observe the effectiveness of such an approach under varying factors, such as cost, computational limitations, and agent experience. Of course, each of these factors will ultimately be constrained by the domain in which the approach is employed. For example, when cost is sufficiently high, agents elect not to communicate, coinciding with the suggestion that in many cases, communication is simply too difficult for ad hoc agents. However, under more favorable conditions, we characterize the considerations under which this style of approach may prove to be a vital mechanism for coordination.

11.3.1 Heuristic Design

Weighted Heuristics

The motivation for weighting heuristics by an approximate state likelihood was to limit the ability for queries at distant future states to supersede states closer to the agent's current decision. From this evaluation, we can see that the weighting of heuristics in many cases significantly improved the performance of the base heuristic, for example as in the case of the local policy entropy heuristic. As a perhaps unexpected result, the weighted random heuristic, while achieving fewer gains than all other weighted heuristics, was nevertheless able to obtain significant results over the baseline. In this tested domain, it is likely that temporally close decisions are more likely to impact one another.

That said, we can imagine a scenario in which such weighting is disadvantageous. Consider a domain in which the agent's current decision depends on the teammate's decision at some arbitrarily distant state. Practically, the heuristic may prune the necessary state from consideration. With this observation, we suggest that the use of a weighting strategy should consider the structure of the domain foremost.

Decision-theoretic vs Information-theoretic

While weighting boosted the overall improvements of information-theoretic heuristics, decision-theoretic heuristics more frequently outperformed the baseline. This is certainly due to omission of utility from consideration in the information-theoretic heuristics, a choice in opposition of the goal of the agent: to maximize its expected utility. However, one information-theoretic heuristic—immediate policy entropy—performed as well as the decision-theoretic heuristics. This is likely due to the high impact of policy uncertainty on the agent's current decision. We

note that the immediate approximate value of information performs equally well and takes into account both uncertainty over teammate policies and the utility associated with each outcome. As it provides a more comprehensive view of the agent's immediate decision, we expect it could match or exceed the immediate policy entropy heuristic in many situations.

Within the set of decision-theoretic heuristics, we distinguish the value of information heuristics from the error-based heuristics, of which the latter achieved fewer significant results. It is worth noting that while similarly related to utility, the two approaches focus on distinct metrics. The error-based heuristics provide a metric related to the variance of utility at stake, i.e. the amount by which the current estimate may be off from the true value. In contrast, the value of information measures the potential for improvement. It is possible for a prediction to result in a high error but a low value of information. As such, the value of information heuristics likely capture more relevant information for the communication process.

11.3.2 Planning Considerations

Greedy Approaches

As in Chapter 9, we are interested in the effectiveness of greedy approaches, as such strategies drastically reduce the computational requirements of a communicative agent. In this evaluation, we observed substantially improved coordination despite small search budgets. This improves upon the results of Chapter 9, though we note substantial differences in the setup. For these experiments, we did not approximate the belief update and policy recomputation. Furthermore, rather than using a rather simplistic learning approach or a manually-authored model-based prior, the agents learned a distribution over teammate policies through past

experience, under the assumption that the agent can perfectly recall every policy it has coordinated with in past episodes. While more accurate, such a capability may not always hold in practice. Additionarlly, while the more accurate computation of the value of information may have been sufficient to boost the success of the greedy approach, we further note substantial domain disparity, as different mazes and populations of teammates were used.

Nevertheless, while greedy approaches may work under certain conditions, consider situations such as the example used in Section 8.1.1 discussing submodularlity. Clearly, the success of the agent can depend on its ability to reason over sets of policy information, necessitating non-greedy search. As common in these considerations, this is a domain-dependent factor.

Search Parameters

Factors such as the cost of communication as well as the accuracy of heuristics in evaluating queries play a significant role in the choice of search parameters. Consider the weighted random evaluator; it performed poorly when only evaluating the heuristic's top choice. However, by increasing the branching factor of the search, the agent's success rate improved substantially. Similarly, with larger sets of information required to coordinate, more iterations will need to be allotted to the search process. In a worst case scenario, successful communication may require a large branching factor and a deep search horizon, in which case this approach may necessitate too much computation to be practically applied.

11.3.3 Domain Considerations

Teammate Policy Variance and Agent Experience

Section 11.2.3 measured the effect of the underlying variance of the population from which teammates are sampled. To our knowledge, no such analysis has been attempted in previous literature[3]. As expected, large variety of teammate populations were correlated with the increase in performance of communicative agents, as agents require further means of reducing uncertainty when observations cannot distinguish policies. Moreover, when relatively inexperienced, communication can compensate for inaccurate priors. Likewise, in scenarios with small populations of teammate strategies, agents require less experience and may not require communication to correctly identify teammate policies.

Domain Structure

In the final experiment of this chapter, we provided an example of how the characteristics of a coordination domain can impact both the need for communication as well as the need for accurately modeling the population of teammate policies. In the maze tested, the coordinating agent was able to coordinate to a high degree of success while possessing no prior experience and not utilizing communication to any significant advantage. To the author's knowledge, analysis of domains has not yet received considerable attention within the context of ad hoc teamwork, though it has been a subject of prior work in coordinated policies in homogeneous multiagent teams [87, 76, 59]. The relationship between the structure of a domain and the need for information poses meaningful problems in the design of coordinating agents as well as of the domains of application.

[3] Albrecht et al. [8] measured the effect of various prior constructions, but focused on the weights of teammate types more so than variance of the population tested.

For example, in an established domain, such as the drop-in robot soccer league [82, 47], the robots are allowed to communicate and have established protocols regarding the message structure and content, typically on the topics of agent locations or agent roles. Currently, these forms of messages are not utilized to any significant benefit, and it is unclear how well teams can and will be able to predict teammate strategies based on such limited information. Ultimately, an analysis of uncertainty in the domain may highlight more granular policy decisions which need to be communicated.

Furthermore, consider the perspective of a someone designing a domain for ad hoc teamwork, potentially a game mixing teams of humans and computer agents. Choices in the design will affect how much information the coordinating agents have access to when working with their human peers. With the results of this chapter in mind, it is clear that such considerations can help reduce the uncertainty over player intentions, reducing the need for explicit communication when observation-based inference is sufficient.

Chapter 12

Discussion

Explicit communication of agent policies is a consideration not to be overlooked in ad hoc teamwork research. Past work often dismisses the possibility of communication, citing the lack of a shared communication protocol. Moreover, in domains where communication is permitted, it is commonly added as a small number of high-level, domain-specific messages (such as the agent's intended role in RoboCup [47]), or it is restricted to sharing only hidden state information not specific to the individual plans of the team members (e.g. bandit information in [10]). This work establishes a communicative strategy for eliciting minimal sets of policy information from teammates, allowing the agent to reason over its own information needs in a well-defined decision problem framework. The advantage of a more granular approach such as ours over broad role communication is the capability to adaptively query varying amounts of policy information as necessary. This permits control over the exactness of beliefs regarding the overall joint plan.

When cooperating with unknown teammates, such communicative agents 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. Agents learning purely through observation require that observations be

informative in order for the agent to correctly infer teammate policies, a property that may not hold in every application. Communication of policy information complements learning agents, as such communicative behavior analyzes the uncertainty within an existing teammate model and advances the information an agent possesses about its teammates, providing an additional means of coordination when learning falls short.

12.1 Contributions

Here, we review the major contributions of this work.

- In Chapter 2, we outlined the decision problem faced by a coordinating agent operating in an ad hoc multiagent team. Notably, this problem is represented as a POMDP, wherein a coordinating agent must use observations of a teammate actions to infer its policy.

- As an example demonstrating the difficulty of ad hoc teamwork, Chapter 5 introduced a modified belief revision approach for the approximation of novel teammate behavior. While the modification achieves significant improvement in the agent's ability to coordinate, it is ultimately an approach that relies on *reacting* to a teammate's behavior, as opposed proactively eliciting information.

- Chapter 6 provided further motivation for policy communication in ad hoc settings. Specifically, we described four conditions under which agents may be uncertain. First, when an agent has yet to learn an accurate distribution over the space of teammate policies, it may omit or underestimate the likelihood of some policies, hindering inference. Second, an agent initially coordinating with teammate my not have gathered enough information to

infer the teammate's policy or to learn a predictive model. As a separate but related consideration, the agent has little control over the informativeness of observations, as they are typically governed by the domain's structure, which may have unavoidable sequences of non-discriminative observations. Finally, when operating in novel coordination problems or new configurations of a coordination state space, policies learned in other settings may not generalize to the constraints of the new environment.

- In Chapter 7, we reexamined the position of a coordinating agent, focusing particularly on an information-centric representation of the agent's beliefs. We described how observations and policy queries combine to form discrete sets of information, over which an agent may evaluate the expected utility gained from transitioning between sets (the value of information). We compose these properties into the Policy Communication Decision Problem, presenting a framework for decision-making when an agent is given the ability to query portions of a teammate's policy. Such a problem is not trivially solved, however, as the information state space is exponentially larger than the state space for the underlying coordination domain. We establish theoretical characteristics of this problem, particularly with regard to its complexity.

- As an initial attempt at tractably computing communication policies, we introduced an approximate, greedy query selection algorithm in Chapter 9. The approach sacrifices accuracy of the predicted value of information and the ability to reason over sequences of queries in exchange for significant computational savings. The results demonstrate the potential for such an approach to identify important decision points to query, yet it was not able to achieve a significant gain in the agent's ability to coordinate.

- To further explore the potential for reducing the evaluation of queries, we proposed pruning the set of queries to be evaluated at each information state. For this purpose, we introduced many heuristic measures to provide pseudo-evaluations, from which we can order queries from most promising to least. With a method of significantly reducing the set of queries, we employ a pruned, sample-based search approach for computing communication policies. We then evaluated the approach, measuring the impacts of search parameterizations, agent experience (accuracy of the initial beliefs), cost of communication, domain structure, and, of course, the choice of heuristic query evaluation. Our results establish the first successful demonstration of agent coordination through policy communication, providing well-characterized evidence for the potential for such approaches.

12.2 Recommendations for Application

Throughout this research, we have overcome many of the technical challenges of reasoning over large sets of information, which has provided many insights regarding the application of heuristics and the considerations involved in making trade-offs between accuracy and computational feasibility. We summarize these lessons as a series of considerations to be made when developing communicative agents for an ad hoc team domains.

- **Analyze the domain of application, assessing the need for communication.** Situations in which an agent observes clear indications of a teammate's intended behavior may not benefit from communication. Furthermore, if the cost of miscoordination is sufficiently low, communication may not be worth the resources involved.

- **Anticipate the variance in teammate policies.** In many cases, a simple domain analysis can gauge the space of teammate policies an agent may encounter. Moreover, if one can predict the relative likelihood of the policies, a hand-crafted prior can often improve an agent's performance before it manages to obtain enough samples to learn a more accurate prior.

- **Identify sets of decisions that must be coordinated together.** Recall that the value of information may not become significant until a set of multiple of decisions is clarified[1]. When the sets are small, greedy approaches may be sufficient. For longer sequences of coordinated policies, a search-based approach may be necessary. This information directly impacts the parameterization of the approach used, e.g. the maximum branching factor and search iterations of a sample-based planner, as evaluated in Chapter 11.

- **Use domain knowledge when selecting or designing a heuristic to evaluate queries.** If coordinated decisions tend to be temporally close, a weighted heuristic may provide an advantage over its unweighted counterpart. Conversely, if an agent's decision depends on a distant choice by a teammate, an unweighted heuristic is better suited for evaluating queries. Utilizing domain knowledge in this way can save substantial computation time by informing the search over communication policies.

12.3 Recommendations for Future Work

In this thesis, we have demonstrated an effective communication procedure that fits the common assumptions of ad hoc teamwork scenario. However, we envision related applications that must operate under alternate conditions, many

[1]Discussed in Section 8.1.1.

of which motivate direct extensions to this work.

12.3.1 Alternative Teammate Policy Representations

Throughout the experiments presented here, we have assumed a teammate's policy is specified over the state space of the coordination problem. In practice, this may not hold, as the teammate may be attempting to reason about its teammates' policies, necessitating a policy over some form of belief representation, much the same way we structure the policy of a coordinating agent. This ultimately results in a much larger space of policy queries to evaluate. Yet, it is not apparent what form the policy should take. The beliefs, as constructed here, are merely probability distributions over potential individual policies. However, other representations can be mapped to such a belief, including past histories of interaction or the information states described in Chapter 7.

12.3.2 Non-stationary Teammate Policies

While the results of this work on communicating policies assumes teammates follow a single policy through the collaborative effort, the work of Chapter 5 as well as [4, 122] motivate considerations for teammates with non-stationary policies. This can occur in two manners. First, if a team is collaborating through multiple successive trials, a teammate may switch policies between trials. Secondly, teammates may switch policies within a trial, deciding to change plans on the fly. In the first case, it is necessary to learn a prior over the possible policies a teammate may choose from. The latter case, however, may entail responsive detection of strategy changes, as used in Chapter 5, and will necessarily require some notion of the assessing the validity of communicated policy information over time.

12.3.3 Other Forms of Policy-Oriented Communication

In order to develop and reason about the implications of communication over policy information, our approach operated on queries of individual state-action pairs. In practice, many other forms of communication can elicit policy information in varying degrees policy coverage. Consider the plan representation used in SharedPlans [57, 58], i.e. a hierarchical task network (HTN). As opposed to a policy, HTNs decompose high-level tasks into smaller tasks, and SharedPlans uses this hierarchical structure when reasoning over the individual intentions of the agents involved. For example, an agent may believe its collaborator can fulfill a task yet the agent may leave uncertain the particular details of how the task is carried out; what matters most is the belief that the task will be successfully completed as part of the larger joint effort. To a degree, our approach can use information from a small number of observations or queries to infer actions across many other states; however, in the absence of correlated actions that permit such inference, it will be helpful to be able to query a teammate for policy information over a subset of states. We note two potential directions for this research, state variable elimination [70], which aims to create abstractions over states by eliminating information that makes unnecessary distinctions between states, and hierarchical solutions to MDPs [62], which focuses on developing macro-actions—local policies defined over a subset of the state space—to solve MDPs.

12.3.4 Learning Communication Strategies

Finally, while we proposed a search-based approach to computing a communication policy, the MDP structure of the decision problem would permit the application of reinforcement learning. It may be possible to use modern deep reinforcement learning techniques, such as [88], though it would require a repre-

sentation of beliefs, states, and actions which can be embedded as inputs to the system, the feasibility of which is domain-specific.

Alternatively, learning can be utilized in more intermediate forms. For the search over information states, we propose heuristics for pruning the space of query *actions*. This mirrors a similar strategy in AlphaGo Zero [128], a hybrid deep learning, policy search approach, in which a policy network provides a policy distribution $\Pr(a \mid s)$ and a value network proves value estimates, $\hat{V}(s)$. The policy distribution is used to bias the search to explore actions that have, in the past, been successful. When the action probabilities are low, they are effectively pruned from consideration. In a similar manner, we can imagine an approach where a heuristic query evaluator is learned, providing $\Pr(q \mid I)$. Furthermore, while we discussed in Chapter 8 the existence of an admissible search heuristic for the communication problem, we noted that such a heuristic often overestimates the value of information, and as a result, we opted to provide the search with no heuristic estimate. Consequently, learning a value estimate for information states is a particularly promising direction for future work.

12.4 The Need for Communicating Ad Hoc Teams Research

In the seminal paper introducing ad hoc teamwork [134], the authors cite the ever-increasing utilization of agents in our world as a motivation for analyzing how agents can learn to interact with others, particularly within collaborative efforts. As such, most existing work focuses on the problem of learning teammate policies and relying on decision-theoretic planning to handle uncertain decisions. Though the work of Albrecht et al. [7] demonstrated the existence of situations

in which observation-based inference can fail, to our knowledge, no other existing work has prescribed alternative mechanisms for coordination in these scenarios. Building on the motivation of [134], we cite the advances in fields such as natural language processing, and human-computer interaction among others that suggest the challenges of communication between heterogeneous agents may not be as impractical as is commonly assumed. In many applications of ad hoc agents, such as the Robocup league or human-computer interaction in the form of a game, restricted forms of communication are frequently used. We believe communication to be an effective—and often necessary—means of coordination, supplementing existing approaches, as we have demonstrated in this thesis. Toward this end, we hope to engage the community in this new research effort!

Appendix A

Selection of α for a Chinese Restaurant Process

The concentration parameter, $\alpha \in \mathbb{N} = \{1, 2, \dots\}$, of a Chinese Restaurant Process [103] is related to the expected likelihood of witnessing a novel observation. Large values of α correspond to high likelihoods of novel observations. Moreover, as the parameter establishes the likelihood of the data observed, we can estimate the parameter using the set of previous observations. Given a set of n observations of teammate policies, $\mu_1, \mu_2, \dots, \mu_n$, the posterior for α is given by

$$\begin{aligned}\Pr(\mu_{1:n} \mid \alpha) &= \Pr(\mu_1 \mid \alpha) \prod_{i=1}^{n-1} \Pr(\mu_{i+1} \mid \mu_{1:i}, \alpha) \\ &\propto \frac{\alpha^m \Gamma(\alpha)}{\Gamma(\alpha + n)}\end{aligned}$$

where m is the number of *unique* observations and Γ is the gamma function. This posterior allows us to use Bayes theorem when combined with a suitable prior over the hyperparameter. For simplicity, we are interested in choosing $\hat{\alpha}$ as the maximum likelihood estimator of the parameter. To do so, we choose a

non-conjugate prior, $\Pr(\alpha) = (1+\alpha)^{-2}$ and, then, estimate the posterior over the parameter as

$$\begin{aligned}\Pr(\alpha \mid \mu_{1:n}) &= \Pr(\mu_{1:n} \mid \alpha) \Pr(\alpha) \\ &\approx \frac{\alpha^m \Gamma(\alpha)}{\Gamma(\alpha+n)(1+\alpha)^2}.\end{aligned}$$

The resulting distribution is unimodal. We utilize hill climbing to find $\hat{\alpha} = \arg\max_{\alpha} \Pr(\alpha \mid \mu_{1:n})$.

Appendix B

Extended Results for Chapter 11

B.1 Communication Search Parameters

Table B.1: Heuristics evaluation with communication branch factor of 1 and 1 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	25	1.000	50.00	50.51	0.001
Mean Absolute Error	50	36	0.923	68.30	46.47	0.113
Mean Squared Error	50	28	0.999	52.50	51.34	0.002
Δ Policy Entropy	50	30	0.996	60.00	49.49	0.015
Approx. Value of Info.	50	50	**0.001***	96.60	2.36	**0.011***
Weighted Action Entropy	50	50	**0.001***	96.40	2.27	**0.012***
Weighted Mean Abs. Error	50	50	**0.001***	97.40	2.52	**0.007***
Weighted Mean Sq. Error	50	50	**0.001***	93.30	4.59	**0.046***
Weighted Δ Policy Ent.	50	50	**0.001***	97.20	2.51	**0.008***
Weighted Approx. VoI	50	50	**0.001***	97.10	2.49	**0.008***
Immediate Policy Ent.	50	50	**0.001***	96.80	2.42	**0.010***
Immediate Value of Info.	50	50	**0.001***	96.80	2.42	**0.010***
Uniform Random	50	33	0.980	66.00	47.85	0.069
State Likelihood	50	50	**0.001***	96.40	2.27	**0.012***
Weighted Uniform Random	50	41	0.602	80.00	37.93	0.795

Table B.2: Heuristics evaluation with communication branch factor of 3 and 1 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	34	0.968	68.00	47.12	0.108
Mean Absolute Error	50	43	0.393	80.60	35.19	0.851
Mean Squared Error	50	33	0.980	62.30	48.90	0.028
Δ Policy Entropy	50	30	0.996	60.00	49.49	0.015
Approx. Value of Info.	50	50	**0.001***	96.70	2.39	**0.010***
Weighted Action Entropy	50	50	**0.001***	96.40	2.27	**0.012***
Weighted Mean Abs. Error	50	50	**0.001***	97.20	2.51	**0.008***
Weighted Mean Sq. Error	50	50	**0.001***	96.60	2.36	**0.011***
Weighted Δ Policy Ent.	50	50	**0.001***	96.50	2.31	**0.011***
Weighted Approx. VoI	50	50	**0.001***	96.80	2.42	**0.010***
Immediate Policy Ent.	50	50	**0.001***	96.10	2.09	**0.013***
Immediate Value of Info.	50	50	**0.001***	96.10	2.09	**0.013***
Uniform Random	50	30	0.996	59.90	49.62	0.015
State Likelihood	50	50	**0.001***	96.50	2.31	**0.011***
Weighted Uniform Random	100	99	**0.000***	96.30	10.12	**0.013***

Table B.3: Heuristics evaluation with communication branch factor of 5 and 1 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	35	0.950	70.00	46.29	0.163
Mean Absolute Error	50	49	**0.008***	90.60	14.45	0.147
Mean Squared Error	50	44	0.288	83.10	33.36	0.880
Δ Policy Entropy	50	31	0.993	62.00	49.03	0.026
Approx. Value of Info.	50	50	**0.001***	96.00	2.02	**0.014***
Weighted Action Entropy	50	50	**0.001***	96.80	2.42	**0.010***
Weighted Mean Abs. Error	50	50	**0.001***	96.70	2.39	**0.010***
Weighted Mean Sq. Error	50	50	**0.001***	96.80	2.99	**0.010***
Weighted Δ Policy Ent.	50	50	**0.001***	96.80	2.42	**0.010***
Weighted Approx. VoI	50	50	**0.001***	96.30	2.22	**0.012***
Immediate Policy Ent.	50	50	**0.001***	96.30	2.22	**0.012***
Immediate Value of Info.	50	50	**0.001***	96.70	2.39	**0.010***
Uniform Random	50	34	0.968	67.50	46.80	0.095
State Likelihood	50	50	**0.001***	96.80	2.42	**0.010***
Weighted Uniform Random	100	97	**0.003***	93.95	16.78	**0.042***

Table B.4: Heuristics evaluation with communication branch factor of 1 and 10 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	35	0.950	70.00	46.29	0.163
Mean Absolute Error	50	36	0.923	68.50	46.59	0.119
Mean Squared Error	50	34	0.968	64.90	48.48	0.055
Δ Policy Entropy	50	37	0.886	74.00	44.31	0.339
Approx. Value of Info.	50	50	**0.001***	97.30	2.52	**0.008***
Weighted Action Entropy	50	50	**0.001***	96.60	2.36	**0.011***
Weighted Mean Abs. Error	50	50	**0.001***	96.40	2.27	**0.012***
Weighted Mean Sq. Error	50	50	**0.001***	94.00	4.95	**0.035***
Weighted Δ Policy Ent.	50	50	**0.001***	96.90	2.45	**0.009***
Weighted Approx. VoI	50	50	**0.001***	97.50	2.53	**0.007***
Immediate Policy Ent.	50	50	**0.001***	96.60	2.36	**0.011***
Immediate Value of Info.	50	50	**0.001***	96.40	2.27	**0.012***
Uniform Random	50	35	0.950	69.70	46.33	0.153
State Likelihood	50	50	**0.001***	93.80	4.80	**0.038***
Weighted Uniform Random	50	44	0.288	84.10	32.62	0.770

Table B.5: Heuristics evaluation with communication branch factor of 3 and 10 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	27	0.999	54.00	50.35	0.002
Mean Absolute Error	50	48	**0.026***	89.90	19.68	0.203
Mean Squared Error	50	48	**0.026***	90.10	20.14	0.194
Δ Policy Entropy	50	29	0.998	58.00	49.86	0.009
Approx. Value of Info.	50	50	**0.001***	96.80	2.42	**0.010***
Weighted Action Entropy	50	50	**0.001***	96.70	3.13	**0.010***
Weighted Mean Abs. Error	50	50	**0.001***	97.20	2.51	**0.008***
Weighted Mean Sq. Error	50	50	**0.001***	97.10	2.49	**0.008***
Weighted Δ Policy Ent.	50	50	**0.001***	97.20	2.51	**0.008***
Weighted Approx. VoI	50	50	**0.001***	96.40	2.27	**0.012***
Immediate Policy Ent.	50	50	**0.001***	96.60	2.36	**0.011***
Immediate Value of Info.	50	50	**0.001***	97.00	2.47	**0.009***
Uniform Random	50	36	0.923	71.70	45.18	0.224
State Likelihood	50	50	**0.001***	96.00	2.02	**0.014***
Weighted Uniform Random	100	100	**0.000***	96.30	3.80	**0.012***

Table B.6: Heuristics evaluation with communication branch factor of 5 and 10 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	30	0.996	60.00	49.49	0.015
Mean Absolute Error	50	49	**0.008***	92.60	14.92	0.076
Mean Squared Error	50	49	**0.008***	92.40	14.99	0.082
Δ Policy Entropy	50	32	0.988	64.00	48.49	0.043
Approx. Value of Info.	50	50	**0.001***	96.60	2.36	**0.011***
Weighted Action Entropy	50	50	**0.001***	96.70	2.39	**0.010***
Weighted Mean Abs. Error	50	50	**0.001***	97.10	2.49	**0.008***
Weighted Mean Sq. Error	50	50	**0.001***	96.90	2.45	**0.009***
Weighted Δ Policy Ent.	50	50	**0.001***	94.10	4.81	**0.033***
Weighted Approx. VoI	50	50	**0.001***	96.90	2.45	**0.009***
Immediate Policy Ent.	50	50	**0.001***	96.60	2.36	**0.011***
Immediate Value of Info.	50	50	**0.001***	93.70	4.72	**0.039***
Uniform Random	50	39	0.773	77.50	42.33	0.581
State Likelihood	50	50	**0.001***	97.00	2.47	**0.009***
Weighted Uniform Random	100	100	**0.000***	96.10	3.14	**0.013***

Table B.7: Heuristics evaluation with communication branch factor of 1 and 20 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	37	0.886	74.00	44.31	0.339
Mean Absolute Error	50	33	0.980	63.30	49.36	0.038
Mean Squared Error	50	32	0.988	61.00	49.64	0.021
Δ Policy Entropy	50	35	0.950	70.00	46.29	0.163
Approx. Value of Info.	50	50	**0.001***	96.60	2.36	**0.011***
Weighted Action Entropy	50	50	**0.001***	96.90	2.45	**0.009***
Weighted Mean Abs. Error	50	50	**0.001***	94.60	5.03	**0.027***
Weighted Mean Sq. Error	50	50	**0.001***	95.80	2.74	**0.015***
Weighted Δ Policy Ent.	50	50	**0.001***	96.80	2.42	**0.010***
Weighted Approx. VoI	50	50	**0.001***	96.60	2.36	**0.011***
Immediate Policy Ent.	50	50	**0.001***	97.30	2.52	**0.008***
Immediate Value of Info.	50	50	**0.001***	96.40	2.27	**0.012***
Uniform Random	50	37	0.886	73.80	44.66	0.330
State Likelihood	50	50	**0.001***	92.60	4.07	0.060
Weighted Uniform Random	50	49	**0.008***	93.80	14.87	**0.049***

Table B.8: Heuristics evaluation with communication branch factor of 3 and 20 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	36	0.923	72.00	45.36	0.239
Mean Absolute Error	50	49	**0.008***	90.70	13.96	0.141
Mean Squared Error	50	41	0.602	77.10	38.85	0.530
Δ Policy Entropy	50	31	0.993	62.00	49.03	0.026
Approx. Value of Info.	50	50	**0.001***	96.30	2.22	**0.012***
Weighted Action Entropy	50	50	**0.001***	97.40	2.52	**0.007***
Weighted Mean Abs. Error	50	50	**0.001***	96.70	2.39	**0.010***
Weighted Mean Sq. Error	50	50	**0.001***	95.90	2.19	**0.015***
Weighted Δ Policy Ent.	50	50	**0.001***	96.30	2.22	**0.012***
Weighted Approx. VoI	50	50	**0.001***	96.80	2.42	**0.010***
Immediate Policy Ent.	50	50	**0.001***	96.70	2.39	**0.010***
Immediate Value of Info.	50	50	**0.001***	93.00	4.52	0.052
Uniform Random	50	36	0.923	71.60	45.35	0.221
State Likelihood	50	50	**0.001***	97.00	2.47	**0.009***
Weighted Uniform Random	100	100	**0.000***	96.00	3.26	**0.014***

Table B.9: Heuristics evaluation with communication branch factor of 5 and 20 iteration(s) per search step.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	41	—	82.00	38.81	—
Action Entropy	50	29	0.998	58.00	49.86	0.009
Mean Absolute Error	50	50	**0.001***	93.40	4.79	**0.044***
Mean Squared Error	100	100	**0.000***	92.15	5.04	0.071
Δ Policy Entropy	100	68	0.980	68.00	46.88	0.055
Approx. Value of Info.	50	50	**0.001***	96.60	2.36	**0.011***
Weighted Action Entropy	50	50	**0.001***	93.60	4.52	**0.041***
Weighted Mean Abs. Error	50	50	**0.001***	96.50	2.31	**0.011***
Weighted Mean Sq. Error	50	50	**0.001***	96.70	2.39	**0.010***
Weighted Δ Policy Ent.	50	50	**0.001***	94.50	4.76	**0.028***
Weighted Approx. VoI	50	50	**0.001***	97.00	2.47	**0.009***
Immediate Policy Ent.	50	50	**0.001***	96.60	2.36	**0.011***
Immediate Value of Info.	50	50	**0.001***	93.90	4.44	**0.036***
Uniform Random	100	72	0.941	71.40	45.35	0.139
State Likelihood	50	50	**0.001***	95.90	1.94	**0.015***
Weighted Uniform Random	100	100	**0.000***	95.75	3.36	**0.016***

B.2 Past Experience

Table B.10: Agent coordinating with 0 episodes of past experience.* denotes significant improvement over baseline.

Heuristic	Experience	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	0	50	28	—	56.00	50.14	—
Action Entropy	0	100	63	0.257	63.00	48.52	0.417
Mean Absolute Error	0	50	30	0.420	60.00	49.49	0.689
Mean Squared Error	0	50	33	0.206	66.00	47.85	0.310
Δ Policy Entropy	0	50	27	0.656	54.00	50.35	0.843
Approx. Value of Info.	0	50	28	0.580	56.00	50.14	1.000
Weighted Action Entropy	0	50	35	0.107	70.00	46.29	0.150
Weighted Mean Abs. Error	0	50	29	0.500	58.00	49.86	0.842
Weighted Mean Sq. Error	0	50	36	0.072	72.00	45.36	0.097
Weighted Δ Policy Ent.	0	50	37	**0.046***	74.00	44.31	0.060
Weighted Approx. Vol	0	50	30	0.420	60.00	49.49	0.689
Immediate Policy Ent.	0	50	31	0.342	62.00	49.03	0.547
Immediate Value of Info.	0	50	31	0.342	62.00	49.03	0.547
Uniform Random	0	50	31	0.342	62.00	49.03	0.547
State Likelihood	0	50	31	0.342	62.00	49.03	0.547
Weighted Uniform Random	0	100	70	0.065	70.00	46.06	0.101

Table B.11: Agent coordinating with 10 episodes of past experience.* denotes significant improvement over baseline.

Heuristic	Experience	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	10	50	29	—	58.00	49.86	—
Action Entropy	10	100	67	0.183	66.70	47.71	0.309
Mean Absolute Error	10	50	50	**0.000***	92.20	5.64	**0.000***
Mean Squared Error	10	50	49	**0.000***	90.80	14.33	**0.000***
Δ Policy Entropy	10	50	34	0.204	67.70	47.15	0.320
Approx. Value of Info.	10	50	50	**0.000***	90.60	6.67	**0.000***
Weighted Action Entropy	10	50	49	**0.000***	92.90	13.71	**0.000***
Weighted Mean Abs. Error	10	50	50	**0.000***	93.30	5.31	**0.000***
Weighted Mean Sq. Error	10	50	50	**0.000***	89.30	9.26	**0.000***
Weighted Δ Policy Ent.	10	50	45	**0.000***	84.80	30.44	**0.002***
Weighted Approx. Vol	10	50	50	**0.000***	90.70	6.47	**0.000***
Immediate Policy Ent.	10	50	41	**0.008***	76.00	39.03	**0.047***
Immediate Value of Info.	10	50	50	**0.000***	91.80	5.87	**0.000***
Uniform Random	10	50	40	**0.015***	75.00	43.69	0.073
State Likelihood	10	50	49	**0.000***	86.20	15.24	**0.000***
Weighted Uniform Random	10	100	93	**0.000***	85.45	25.44	**0.001***

Table B.12: Agent coordinating with 100 episodes of past experience.* denotes significant improvement over baseline.

Heuristic	Experience	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	50	40	—	80.00	40.41	—
Action Entropy	100	100	74	0.845	74.00	44.08	0.408
Mean Absolute Error	100	50	50	**0.001***	94.20	4.99	**0.017***
Mean Squared Error	100	50	50	**0.001***	92.90	5.35	**0.030***
Δ Policy Entropy	100	50	32	0.978	64.00	48.49	0.076
Approx. Value of Info.	100	50	50	**0.001***	96.80	2.42	**0.005***
Weighted Action Entropy	100	50	50	**0.001***	96.90	2.45	**0.005***
Weighted Mean Abs. Error	100	50	50	**0.001***	96.60	2.36	**0.006***
Weighted Mean Sq. Error	100	50	50	**0.001***	97.00	2.47	**0.005***
Weighted Δ Policy Ent.	100	50	50	**0.001***	96.70	2.39	**0.005***
Weighted Approx. Vol	100	50	50	**0.001***	96.70	2.39	**0.005***
Immediate Policy Ent.	100	50	50	**0.001***	96.80	2.42	**0.005***
Immediate Value of Info.	100	50	50	**0.001***	94.70	4.45	**0.014***
Uniform Random	100	50	32	0.978	62.70	48.44	0.055
State Likelihood	100	50	50	**0.001***	96.80	2.42	**0.005***
Weighted Uniform Random	100	100	100	**0.000***	96.40	3.18	**0.006***

Table B.13: Agent coordinating with 1000 episodes of past experience.* denotes significant improvement over baseline.

Heuristic	Experience	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	1000	50	34	—	68.00	47.12	—
Action Entropy	1000	100	71	0.422	71.00	45.60	0.711
Mean Absolute Error	1000	50	34	0.585	68.00	47.12	1.000
Mean Squared Error	1000	50	31	0.799	62.00	49.03	0.534
Δ Policy Entropy	1000	100	69	0.522	69.00	46.48	0.902
Approx. Value of Info.	1000	50	50	**0.000***	96.40	2.27	**0.000***
Weighted Action Entropy	1000	50	50	**0.000***	96.60	2.36	**0.000***
Weighted Mean Abs. Error	1000	50	50	**0.000***	96.50	2.31	**0.000***
Weighted Mean Sq. Error	1000	50	32	0.737	64.00	48.49	0.677
Weighted Δ Policy Ent.	1000	50	50	**0.000***	96.50	2.31	**0.000***
Weighted Approx. Vol	1000	50	50	**0.000***	96.60	2.36	**0.000***
Immediate Policy Ent.	1000	50	50	**0.000***	96.70	2.39	**0.000***
Immediate Value of Info.	1000	50	50	**0.000***	96.50	2.31	**0.000***
Uniform Random	1000	50	33	0.665	66.00	47.85	0.834
State Likelihood	1000	50	50	**0.000***	96.70	2.39	**0.000***
Weighted Uniform Random	1000	50	48	**0.000***	93.70	19.48	**0.001***

B.3 Population Dynamics

Table B.14: Agent coordinating with 10 experience with 5 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	87	—	87.00	33.80	—
Action Entropy	150	133	0.418	88.10	31.80	0.797
Mean Absolute Error	50	50	0.004	94.80	7.28	**0.029***
Mean Squared Error	50	50	0.004	92.00	10.64	0.179
Δ Policy Entropy	50	45	0.404	87.60	30.54	0.913
Approx. Value of Info.	50	50	0.004	91.90	13.09	0.206
Weighted Action Entropy	50	46	0.267	86.90	29.84	0.985
Weighted Mean Abs. Error	50	50	0.004	92.30	10.46	0.153
Weighted Mean Sq. Error	50	50	0.004	91.50	11.35	0.231
Weighted Δ Policy Ent.	100	96	**0.020***	91.90	20.25	0.215
Weighted Approx. Vol	50	50	0.004	94.30	5.98	**0.038***
Immediate Policy Ent.	88	88	**0.000***	83.24	29.07	0.413
Immediate Value of Info.	50	50	0.004	94.50	12.75	0.052
Uniform Random	50	48	0.069	89.50	22.75	0.593
State Likelihood	50	50	0.004	91.70	11.50	0.212
Weighted Uniform Random	100	98	**0.003***	93.10	14.72	0.100

Table B.15: Agent coordinating with 100 experience with 5 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	91	—	91.00	28.76	—
Action Entropy	150	140	0.327	92.10	25.16	0.756
Mean Absolute Error	50	50	0.023	94.70	6.42	0.222
Mean Squared Error	50	49	0.096	90.90	15.90	0.978
Δ Policy Entropy	50	46	0.552	90.80	27.56	0.967
Approx. Value of Info.	50	50	0.023	90.60	14.94	0.911
Weighted Action Entropy	100	96	0.125	92.60	20.36	0.650
Weighted Mean Abs. Error	50	50	0.023	90.90	13.16	0.977
Weighted Mean Sq. Error	50	50	0.023	87.30	15.88	0.312
Weighted Δ Policy Ent.	50	45	0.697	87.50	31.25	0.509
Weighted Approx. Vol	50	50	0.023	88.80	18.37	0.571
Immediate Policy Ent.	50	50	0.023	79.60	30.60	0.031
Immediate Value of Info.	50	50	0.023	89.70	18.94	0.741
Uniform Random	100	93	0.398	79.95	34.39	0.015
State Likelihood	50	50	0.023	91.10	7.58	0.974
Weighted Uniform Random	100	99	**0.009***	89.45	15.21	0.634

Table B.16: Agent coordinating with 1000 experience with 5 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	88	—	88.00	32.66	—
Action Entropy	150	135	0.382	89.27	30.28	0.757
Mean Absolute Error	50	50	0.006	95.00	5.25	**0.039***
Mean Squared Error	50	50	0.006	96.10	3.82	**0.016***
Δ Policy Entropy	50	45	0.473	89.20	30.09	0.823
Approx. Value of Info.	50	50	0.006	88.70	15.81	0.860
Weighted Action Entropy	50	45	0.473	86.60	30.65	0.797
Weighted Mean Abs. Error	50	50	0.006	93.50	7.91	0.114
Weighted Mean Sq. Error	50	50	0.006	88.60	13.44	0.874
Weighted Δ Policy Ent.	100	98	**0.005***	93.85	16.05	0.110
Weighted Approx. Vol	50	50	0.006	93.40	9.87	0.131
Immediate Policy Ent.	50	50	0.006	85.40	23.30	0.576
Immediate Value of Info.	50	50	0.006	85.20	22.38	0.539
Uniform Random	50	49	0.033	85.40	23.86	0.581
State Likelihood	50	50	0.006	92.60	6.72	0.179
Weighted Uniform Random	100	98	**0.005***	91.30	18.54	0.381

Table B.17: Agent coordinating with 10 experience with 25 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	68	—	68.00	46.88	—
Action Entropy	150	104	0.465	68.93	46.17	0.877
Mean Absolute Error	50	50	0.000	92.60	4.43	**0.000***
Mean Squared Error	50	50	0.000	92.20	6.79	**0.000***
Δ Policy Entropy	50	34	0.576	67.20	47.25	0.922
Approx. Value of Info.	50	49	0.000	89.90	14.12	**0.000***
Weighted Action Entropy	50	47	0.000	88.10	23.05	**0.001***
Weighted Mean Abs. Error	50	50	0.000	87.30	15.29	**0.000***
Weighted Mean Sq. Error	50	49	0.000	84.90	18.53	**0.002***
Weighted Δ Policy Ent.	50	49	0.000	92.70	13.75	**0.000***
Weighted Approx. VoI	50	49	0.000	89.20	14.01	**0.000***
Immediate Policy Ent.	50	42	0.027	77.60	37.20	0.176
Immediate Value of Info.	50	50	0.000	89.50	8.41	**0.000***
Uniform Random	100	68	0.560	63.35	48.11	0.490
State Likelihood	50	50	0.000	85.80	8.65	**0.000***
Weighted Uniform Random	100	93	**0.000***	86.45	25.75	**0.001***

Table B.18: Agent coordinating with 100 experience with 25 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	78	—	78.00	41.63	—
Action Entropy	150	115	0.654	76.57	42.39	0.791
Mean Absolute Error	50	50	0.000	81.70	14.34	0.426
Mean Squared Error	50	50	0.000	83.10	12.33	0.261
Δ Policy Entropy	50	38	0.689	74.30	45.04	0.628
Approx. Value of Info.	100	100	**0.000***	90.15	7.26	**0.005***
Weighted Action Entropy	50	48	0.003	92.00	19.69	**0.006***
Weighted Mean Abs. Error	50	50	0.000	82.60	14.08	0.321
Weighted Mean Sq. Error	50	50	0.000	82.80	13.22	0.295
Weighted Δ Policy Ent.	50	48	0.003	92.90	19.82	**0.003***
Weighted Approx. VoI	50	50	0.000	91.20	5.30	**0.002***
Immediate Policy Ent.	110	106	**0.000***	82.23	20.89	0.361
Immediate Value of Info.	50	50	0.000	92.30	7.71	**0.001***
Uniform Random	60	48	0.464	65.00	47.10	0.080
State Likelihood	50	50	0.000	90.80	6.01	**0.003***
Weighted Uniform Random	100	96	**0.000***	85.10	22.28	0.135

Table B.19: Agent coordinating with 1000 experience with 25 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	78	—	78.00	41.63	—
Action Entropy	150	119	0.460	78.60	40.79	0.911
Mean Absolute Error	50	50	0.000	83.90	10.61	0.185
Mean Squared Error	50	50	0.000	85.30	8.89	0.096
Δ Policy Entropy	50	40	0.477	78.40	40.70	0.955
Approx. Value of Info.	50	50	0.000	88.40	9.82	**0.019***
Weighted Action Entropy	50	46	0.024	87.30	27.35	0.104
Weighted Mean Abs. Error	50	50	0.000	82.40	13.06	0.336
Weighted Mean Sq. Error	50	50	0.000	87.50	8.28	**0.030***
Weighted Δ Policy Ent.	50	45	0.054	88.00	29.73	0.093
Weighted Approx. VoI	50	50	0.000	91.30	5.70	**0.002***
Immediate Policy Ent.	50	50	0.000	82.20	14.11	0.365
Immediate Value of Info.	50	50	0.000	93.70	3.16	**0.000***
Uniform Random	60	50	0.273	68.75	44.14	0.192
State Likelihood	50	50	0.000	89.40	7.05	**0.009***
Weighted Uniform Random	150	149	**0.000***	84.50	15.00	0.137

Table B.20: Agent coordinating with 10 experience with 125 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	73	—	73.00	44.62	—
Action Entropy	150	108	0.623	71.40	45.41	0.783
Mean Absolute Error	50	50	0.000	90.60	7.60	**0.000***
Mean Squared Error	50	50	0.000	91.90	7.42	**0.000***
Δ Policy Entropy	50	35	0.721	68.50	47.62	0.579
Approx. Value of Info.	50	49	0.000	89.50	15.66	**0.001***
Weighted Action Entropy	50	49	0.000	92.40	13.71	**0.000***
Weighted Mean Abs. Error	50	50	0.000	82.10	16.94	0.074
Weighted Mean Sq. Error	50	50	0.000	88.00	14.07	**0.003***
Weighted Δ Policy Ent.	50	46	0.004	87.80	26.27	**0.012***
Weighted Approx. VoI	50	50	0.000	90.90	5.41	**0.000***
Immediate Policy Ent.	50	37	0.530	66.80	44.02	0.420
Immediate Value of Info.	50	50	0.000	90.30	7.59	**0.000***
Uniform Random	50	42	0.096	78.40	37.57	0.438
State Likelihood	50	50	0.000	88.40	6.88	**0.001***
Weighted Uniform Random	100	96	**0.000***	89.55	20.21	**0.001***

Table B.21: Agent coordinating with 100 experience with 125 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	56	—	56.00	49.89	—
Action Entropy	150	85	0.510	56.67	49.72	0.918
Mean Absolute Error	50	50	0.000	93.30	4.70	**0.000***
Mean Squared Error	50	50	0.000	92.50	5.46	**0.000***
Δ Policy Entropy	50	34	0.107	68.00	47.12	0.152
Approx. Value of Info.	50	50	0.000	96.40	2.27	**0.000***
Weighted Action Entropy	50	50	0.000	94.20	4.78	**0.000***
Weighted Mean Abs. Error	50	50	0.000	96.30	2.22	**0.000***
Weighted Mean Sq. Error	50	50	0.000	97.20	2.51	**0.000***
Weighted Δ Policy Ent.	50	50	0.000	96.80	2.42	**0.000***
Weighted Approx. VoI	50	50	0.000	96.00	2.02	**0.000***
Immediate Policy Ent.	50	50	0.000	96.40	2.27	**0.000***
Immediate Value of Info.	50	50	0.000	94.30	4.74	**0.000***
Uniform Random	50	35	0.069	69.10	46.41	0.115
State Likelihood	50	50	0.000	96.30	2.22	**0.000***
Weighted Uniform Random	100	100	**0.000***	95.20	3.69	**0.000***

Table B.22: Agent coordinating with 1000 experience with 125 maximum unique teammate policies.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	100	74	—	74.00	44.08	—
Action Entropy	150	107	0.727	71.27	45.33	0.635
Mean Absolute Error	50	50	0.000	80.70	11.43	0.156
Mean Squared Error	100	100	**0.000***	82.85	9.78	0.053
Δ Policy Entropy	50	36	0.679	71.90	45.52	0.788
Approx. Value of Info.	50	50	0.000	92.80	3.52	**0.000***
Weighted Action Entropy	100	89	**0.005***	86.15	30.74	**0.025***
Weighted Mean Abs. Error	100	100	**0.000***	83.45	11.07	**0.040***
Weighted Mean Sq. Error	50	50	0.000	82.70	10.98	0.065
Weighted Δ Policy Ent.	50	44	0.036	85.90	32.76	0.065
Weighted Approx. VoI	50	50	0.000	91.80	5.23	**0.000***
Immediate Policy Ent.	100	92	**0.001***	81.05	27.20	0.175
Immediate Value of Info.	50	50	0.000	95.00	0.00	**0.000***
Uniform Random	100	79	0.252	65.70	42.23	0.176
State Likelihood	50	50	0.000	91.40	4.74	**0.000***
Weighted Uniform Random	50	49	0.000	84.50	16.48	**0.037***

B.4 Communication Cost

Table B.23: Agent coordinating with communication cost $C(q) = 1.$* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	37	—	74.00	44.31	—
Action Entropy	50	30	0.956	60.00	49.49	0.139
Mean Absolute Error	50	50	**0.000***	98.66	1.04	**0.000***
Mean Squared Error	50	50	**0.000***	98.72	1.29	**0.000***
Δ Policy Entropy	50	33	0.862	66.00	47.85	0.388
Approx. Value of Info.	50	50	**0.000***	96.90	1.74	**0.001***
Weighted Action Entropy	50	50	**0.000***	98.98	0.14	**0.000***
Weighted Mean Abs. Error	100	100	**0.000***	97.04	2.21	**0.001***
Weighted Mean Sq. Error	50	50	**0.000***	98.58	1.05	**0.000***
Weighted Δ Policy Ent.	50	50	**0.000***	98.14	0.35	**0.000***
Weighted Approx. VoI	50	50	**0.000***	95.90	1.88	**0.001***
Immediate Policy Ent.	50	50	**0.000***	99.36	0.48	**0.000***
Immediate Value of Info.	50	50	**0.000***	93.44	4.34	**0.003***
Uniform Random	50	39	0.408	76.54	41.97	0.769
State Likelihood	50	50	**0.000***	97.38	1.43	**0.000***
Weighted Uniform Random	50	50	**0.000***	97.74	1.77	**0.000***

Table B.24: Agent coordinating with communication cost $C(q) = 5$.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	37	—	74.00	44.31	—
Action Entropy	100	61	0.962	61.00	49.02	0.105
Mean Absolute Error	50	50	**0.000***	92.00	4.04	**0.006***
Mean Squared Error	50	50	**0.000***	92.00	5.15	**0.006***
Δ Policy Entropy	50	36	0.674	72.00	45.36	0.824
Approx. Value of Info.	50	50	**0.000***	96.40	2.27	**0.001***
Weighted Action Entropy	50	50	**0.000***	93.80	4.58	**0.003***
Weighted Mean Abs. Error	50	50	**0.000***	96.60	2.36	**0.001***
Weighted Mean Sq. Error	50	50	**0.000***	97.00	2.47	**0.001***
Weighted Δ Policy Ent.	50	50	**0.000***	95.00	4.95	**0.002***
Weighted Approx. VoI	50	50	**0.000***	96.30	2.22	**0.001***
Immediate Policy Ent.	50	50	**0.000***	96.50	2.31	**0.001***
Immediate Value of Info.	50	50	**0.000***	94.80	4.84	**0.002***
Uniform Random	50	34	0.811	66.90	46.85	0.438
State Likelihood	50	50	**0.000***	96.80	2.42	**0.001***
Weighted Uniform Random	50	50	**0.000***	95.30	4.09	**0.001***

Table B.25: Agent coordinating with communication cost $C(q) = 10$.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	37	—	74.00	44.31	—
Action Entropy	50	33	0.862	66.00	47.85	0.388
Mean Absolute Error	50	34	0.811	64.80	45.46	0.308
Mean Squared Error	50	30	0.956	59.20	48.98	0.116
Δ Policy Entropy	50	30	0.956	60.00	49.49	0.139
Approx. Value of Info.	50	50	**0.000***	94.00	4.95	**0.003***
Weighted Action Entropy	50	50	**0.000***	93.60	4.85	**0.003***
Weighted Mean Abs. Error	50	50	**0.000***	94.00	4.95	**0.003***
Weighted Mean Sq. Error	100	99	**0.000***	92.90	10.66	**0.004***
Weighted Δ Policy Ent.	50	50	**0.000***	93.80	4.90	**0.003***
Weighted Approx. VoI	50	50	**0.000***	92.80	4.54	**0.004***
Immediate Policy Ent.	100	100	**0.000***	93.20	4.69	**0.004***
Immediate Value of Info.	50	50	**0.000***	94.20	4.99	**0.002***
Uniform Random	50	33	0.862	66.00	47.85	0.388
State Likelihood	50	50	**0.000***	93.40	4.79	**0.003***
Weighted Uniform Random	50	47	**0.006***	89.20	23.37	**0.035***

Table B.26: Agent coordinating with communication cost $C(q) = 99$.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	37	—	74.00	44.31	—
Action Entropy	50	39	0.408	78.00	41.85	0.644
Mean Absolute Error	50	35	0.748	70.00	46.29	0.660
Mean Squared Error	50	33	0.862	66.00	47.85	0.388
Δ Policy Entropy	50	39	0.408	78.00	41.85	0.644
Approx. Value of Info.	50	33	0.862	66.00	47.85	0.388
Weighted Action Entropy	50	29	0.972	58.00	49.86	0.093
Weighted Mean Abs. Error	50	35	0.748	70.00	46.29	0.660
Weighted Mean Sq. Error	50	36	0.674	72.00	45.36	0.824
Weighted Δ Policy Ent.	50	35	0.748	70.00	46.29	0.660
Weighted Approx. Vol	50	30	0.956	60.00	49.49	0.139
Immediate Policy Ent.	50	34	0.811	68.00	47.12	0.513
Immediate Value of Info.	50	30	0.956	60.00	49.49	0.139
Uniform Random	50	31	0.934	62.00	49.03	0.202
State Likelihood	50	32	0.903	64.00	48.49	0.284
Weighted Uniform Random	100	73	0.624	73.00	44.62	0.897

B.5 Domain Structure

Table B.27: Agent coordinating with 0 past episodes of experience.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	48	—	96.00	19.79	—
Action Entropy	150	144	0.638	95.60	19.64	0.902
Mean Absolute Error	50	43	0.985	84.30	35.08	0.043
Mean Squared Error	50	45	0.944	88.60	30.61	0.155
Δ Policy Entropy	100	89	0.967	88.05	31.21	0.060
Approx. Value of Info.	50	48	0.691	94.90	19.86	0.782
Weighted Action Entropy	50	48	0.691	95.00	19.92	0.802
Weighted Mean Abs. Error	100	94	0.812	91.90	23.67	0.266
Weighted Mean Sq. Error	50	44	0.970	85.80	32.31	0.061
Weighted Δ Policy Ent.	50	46	0.898	89.60	27.10	0.181
Weighted Approx. Vol	50	48	0.691	94.20	20.11	0.653
Immediate Policy Ent.	50	47	0.819	93.50	23.93	0.571
Immediate Value of Info.	50	45	0.944	87.20	29.92	0.086
Uniform Random	50	46	0.898	91.40	27.29	0.337
State Likelihood	50	47	0.819	91.10	23.95	0.268
Weighted Uniform Random	100	87	0.985	85.60	33.50	0.019

Table B.28: Agent coordinating with 10 past episodes of experience.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	46	—	92.00	27.40	—
Action Entropy	100	91	0.687	90.45	28.66	0.748
Mean Absolute Error	50	42	0.939	82.40	36.98	0.144
Mean Squared Error	50	45	0.757	88.80	30.35	0.581
Δ Policy Entropy	50	45	0.757	88.10	30.59	0.504
Approx. Value of Info.	50	47	0.500	90.40	23.62	0.755
Weighted Action Entropy	50	49	0.181	94.50	14.99	0.573
Weighted Mean Abs. Error	50	47	0.500	90.60	24.17	0.787
Weighted Mean Sq. Error	100	95	0.347	91.45	22.31	0.902
Weighted Δ Policy Ent.	50	46	0.643	88.00	28.03	0.472
Weighted Approx. Vol	100	95	0.347	88.45	23.78	0.437
Immediate Policy Ent.	50	47	0.500	91.30	24.53	0.893
Immediate Value of Info.	50	47	0.500	80.60	26.10	0.036
Uniform Random	50	47	0.500	91.70	23.79	0.954
State Likelihood	50	48	0.339	65.90	24.92	0.000
Weighted Uniform Random	100	93	0.531	89.70	25.54	0.621

Table B.29: Agent coordinating with 100 past episodes of experience.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	47	—	94.00	23.99	—
Action Entropy	100	91	0.830	90.25	28.59	0.400
Mean Absolute Error	50	48	0.500	89.90	19.68	0.353
Mean Squared Error	50	49	0.309	90.00	14.95	0.320
Δ Policy Entropy	50	44	0.920	85.40	32.46	0.135
Approx. Value of Info.	50	48	0.500	88.30	19.45	0.195
Weighted Action Entropy	50	46	0.782	87.20	27.16	0.188
Weighted Mean Abs. Error	50	47	0.661	81.60	24.46	0.012
Weighted Mean Sq. Error	50	43	0.954	76.80	34.07	0.004
Weighted Δ Policy Ent.	50	48	0.500	89.80	19.82	0.342
Weighted Approx. Vol	50	50	0.121	90.10	7.18	0.275
Immediate Policy Ent.	50	46	0.782	89.70	27.00	0.402
Immediate Value of Info.	50	50	0.121	84.60	11.33	0.015
Uniform Random	50	45	0.866	83.90	31.16	0.073
State Likelihood	50	48	0.500	85.40	19.79	0.053
Weighted Uniform Random	100	96	0.429	89.85	20.08	0.296

Table B.30: Agent coordinating with 1000 past episodes of experience.* denotes significant improvement over baseline.

Heuristic	Trials	Successes	$p_{success}$	Reward Avg.	Reward Std.	p_{util}
No Comm.	50	45	—	90.00	30.30	—
Action Entropy	100	86	0.829	85.15	34.92	0.382
Mean Absolute Error	100	91	0.528	87.70	28.05	0.654
Mean Squared Error	100	91	0.528	87.35	27.89	0.606
Δ Policy Entropy	50	45	0.630	87.00	30.77	0.624
Approx. Value of Info.	100	92	0.448	87.05	26.25	0.559
Weighted Action Entropy	50	45	0.630	82.20	29.59	0.196
Weighted Mean Abs. Error	50	46	0.500	76.20	26.43	0.017
Weighted Mean Sq. Error	100	95	0.206	85.00	21.81	0.302
Weighted Δ Policy Ent.	50	47	0.357	90.50	24.44	0.928
Weighted Approx. Vol	50	48	0.218	89.70	19.81	0.953
Immediate Policy Ent.	50	45	0.630	87.40	30.64	0.671
Immediate Value of Info.	50	47	0.357	87.60	23.78	0.661
Uniform Random	50	48	0.218	92.70	21.67	0.610
State Likelihood	50	48	0.218	93.20	19.94	0.534
Weighted Uniform Random	100	90	0.604	83.25	30.37	0.202

Appendix C

MDP Notation Reference

Decision Problem	State Space	Actions	Transition	Reward	Observations	Observation Likelihoods
Multiagent MDP	$s \in S$	$\bar{a} \in \bar{A}$	$T(s, \bar{a}, s')$	$R(s, \bar{a}, s')$	—	—
Individual Agent in MMDP	$\tilde{s} \in \tilde{S}$	$a_i \in A_i$	$\tilde{T}(\tilde{s}, a_0, \tilde{s}')$	$\tilde{R}(\tilde{s}, a_0, \tilde{s}')$	—	—
Individual Agent in PO-MMDP	$\tilde{s} \in \tilde{S}$	$a_i \in A_i$	$\tilde{T}(\tilde{s}, a_0, \tilde{s}')$	$\tilde{R}(\tilde{s}, a_0, \tilde{s}')$	$\mathsf{o} \in \mathsf{O}$	$\Omega(\mathsf{o} \mid \tilde{s}, a_0, \tilde{s}')$
Belief MDP	$\mathsf{b} \in \mathsf{S_b}$	$a_i \in A_i$	$\mathsf{T}(\mathsf{b}, a_0, \mathsf{b}')$	$\mathsf{R}(\mathsf{b}, a_0, \mathsf{b}')$	—	—
Communication MDP	$I \in \mathcal{I}$	$q \in \mathcal{Q}$	$\mathcal{M}(I, q, I')$	$\mathcal{U}(I, q, I')$	—	—

Table C.1: Summary of notation used for various decision problems.

Bibliography

[1] Aswin Thomas Abraham and Kevin McGee. Ai for dynamic team-mate adaptation in games. In *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*, pages 419–426. IEEE, 2010.

[2] Noa Agmon, Samuel Barrett, and Peter Stone. Modeling uncertainty in leading ad hoc teams. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 397–404. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[3] Noa Agmon and Peter Stone. 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*, pages 341–348. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[4] Stefano V Albrecht, Jacob W Crandall, and Subramanian Ramamoorthy. Belief and truth in hypothesised behaviours. *arXiv preprint arXiv:1507.07688*, 2015.

[5] Stefano V Albrecht and Subramanian Ramamoorthy. 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*, pages 349–356. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[6] Stefano V Albrecht and Subramanian Ramamoorthy. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1155–1156. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[7] Stefano V Albrecht and Subramanian Ramamoorthy. 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*, pages 12–21, 2014.

[8] Stefano Vittorino Albrecht, Jacob William Crandall, and Subramanian Ramamoorthy. An empirical study on the practical impact of prior beliefs over policy types. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1988–1994, 2015.

[9] Karl J Åström. Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

[10] Samuel Barrett, Noa Agmon, Noam Hazon, Sarit Kraus, and Peter Stone. Communicating with unknown teammates. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1433–1434. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[11] Samuel Barrett and Peter Stone. Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork. In *AAAI*, pages 2010–2016, 2015.

[12] Samuel Barrett, Peter Stone, and Sarit Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 567–574. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

[13] Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Learning teammate models for ad hoc teamwork. In *AAMAS Adaptive Learning Agents (ALA) Workshop*, 2012.

[14] Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Teamwork with limited knowledge of teammates. In *AAAI*, 2013.

[15] Jennifer L Barry. *Fast approximate hierarchical solution of MDPs*. PhD thesis, Massachusetts Institute of Technology, 2009.

[16] Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[17] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[18] Richard Bellman. A markovian decision process. Technical report, DTIC Document, 1957.

[19] Miroslav Benda. On optimal cooperation of knowledge sources. *Technical Report BCS-G2010-28*, 1986.

[20] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.

[21] Daniel S Bernstein, Shlomo Zilberstein, and Neil Immerman. The complexity of decentralized control of markov decision processes. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 32–37. Morgan Kaufmann Publishers Inc., 2000.

[22] Mustafa Bilgic and Lise Getoor. Active inference for collective classification. In *AAAI*, 2010.

[23] Avrim Blum and Yishay Monsour. Learning, regret minimization, and equilibria. In E. Tardos N. Nisan, T. Roughgarden and V. Vazirani, editors, *Algorithmic Game Theory*. Cambridge University Press, 2007.

[24] Olivier Bousquet and Manfred K Warmuth. Tracking a small set of experts by mixing past posteriors. *The Journal of Machine Learning Research*, 3:363–396, 2003.

[25] Craig Boutilier. Planning, learning and coordination in multiagent decision processes. In *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*, pages 195–210. Morgan Kaufmann Publishers Inc., 1996.

[26] Michael Bratman. Two faces of intention. *The Philosophical Review*, pages 375–405, 1984.

[27] Michael Bratman. *Intention, plans, and practical reason*. Harvard University Press, 1987.

[28] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *Computational Intelligence and AI in Games, IEEE Transactions on*, 4(1):1–43, 2012.

[29] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, And Cybernetics-Part C: Applications and Reviews, 38 (2), 2008*, 2008.

[30] David Carmel and Shaul Markovitch. Pruning algorithms for multi-model adversary search. *Artificial Intelligence*, 99(2):325–355, 1998.

[31] Doran Chakraborty and Peter Stone. Cooperating with a markovian ad hoc teammate. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1085–1092. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[32] Georgios Chalkiadakis and Craig Boutilier. Coordination in multiagent reinforcement learning: A bayesian approach. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 709–716. ACM, 2003.

[33] Muthukumaran Chandrasekaran, Prashant Doshi, Yifeng Zeng, and Yingke Chen. Team behavior in interactive dynamic influence diagrams with applications to ad hoc teams. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1559–1560. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[34] Guillaume Chaslot, Sander Bakkes, Istvan Szita, and Pieter Spronck. Monte-carlo tree search: A new framework for game ai. In *Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment International Conference (AIIDE 2008)*, 2008.

[35] Guillaume MJ-B Chaslot, Mark HM Winands, and H Jaap van Den Herik. Parallel monte-carlo tree search. In *Computers and Games*, pages 60–71. Springer, 2008.

[36] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998:746–752, 1998.

[37] Philip R Cohen and Hector J Levesque. Persistence, intention, and commitment. *Reasoning about actions and plans*, pages 297–340, 1990.

[38] Philip R Cohen and Hector J Levesque. Teamwork. *Nous*, pages 487–512, 1991.

[39] Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. Building explainable artificial intelligence systems. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1766. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[40] Peter I Cowling, Edward J Powley, and Daniel Whitehouse. Information set monte carlo tree search. *Computational Intelligence and AI in Games, IEEE Transactions on*, 4(2):120–143, 2012.

[41] Maria Cutumisu and Duane Szafron. An architecture for game behavior ai: Behavior multi-queues. In *AIIDE*, 2009.

[42] Prashant Doshi, Yifeng Zeng, and Qiongyu Chen. Graphical models for interactive pomdps: representations and solutions. *Autonomous Agents and Multi-Agent Systems*, 18(3):376–416, 2009.

[43] Edmund H Durfee. Blissful ignorance: Knowing just enough to coordinate well. *Ann Arbor*, 148109:2110, 1995.

[44] Markus Enzenberger, Martin Muller, Broderick Arneson, and Richard Segal. FuegoâĂŤan open-source framework for board games and go engine based on monte carlo tree search. *Computational Intelligence and AI in Games, IEEE Transactions on*, 2(4):259–270, 2010.

[45] Hilmar Finnsson and Yngvi Björnsson. Simulation-based approach to general game playing. In *AAAI*, volume 8, pages 259–264, 2008.

[46] Katie Genter, Noa Agmon, and Peter Stone. Ad hoc teamwork for leading a flock. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 531–538. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[47] Katie Genter, Tim Laue, and Peter Stone. The robocup 2014 spl drop-in player competition: Experiments in teamwork without pre-coordination. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1745–1746. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[48] Katie Genter, Shun Zhang, and Peter Stone. Determining placements of influencing agents in a flock. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 247–255. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[49] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *International Workshop on Agent Theories, Architectures, and Languages*, pages 1–10. Springer, 1998.

[50] Itzhak Gilboa. The complexity of computing best-response automata in repeated games. *Journal of economic theory*, 45(2):342–352, 1988.

[51] Piotr J Gmytrasiewicz and Edmund H Durfee. Toward a theory of honesty and trust among communicating autonomous agents. *Group Decision and Negotiation*, 2(3):237–258, 1993.

[52] Piotr J Gmytrasiewicz and Edmund H Durfee. A rigorous, operational formalization of recursive modeling. In *ICMAS*, pages 125–132, 1995.

[53] Piotr J Gmytrasiewicz, Edmund H Durfee, and Jeffrey Rosenschein. Toward rational communicative behavior. In *AAAI Fall Symposium on Embodied Language*, pages 35–43, 1995.

[54] Piotr J Gmytrasiewicz, Edmund H Durfee, and David K Wehe. A decision-theoretic approach to coordinating multi-agent interactions. In *IJCAI*, volume 91, pages 63–68, 1991.

[55] Piotr J Gmytrasiewicz, Edmund H Durfee, and David K Wehe. The utility of communication in coordinating intelligent agents. In *AAAI*, pages 166–172, 1991.

[56] Claudia V Goldman and Shlomo Zilberstein. Optimizing information exchange in cooperative multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 137–144. ACM, 2003.

[57] Barbara J Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[58] Barbara J Grosz and Sarit Kraus. The evolution of sharedplans. In *Foundations of rational agency*, pages 227–262. Springer, 1999.

[59] Carlos Guestrin, Shobha Venkataraman, and Daphne Koller. Context-specific multiagent coordination and planning with factored mdps. In *AAAI/IAAI*, pages 253–259, 2002.

[60] Eric A Hansen and Shlomo Zilberstein. LaoâĹŮ: A heuristic search algorithm that finds solutions with loops. *Artificial Intelligence*, 129(1-2):35–62, 2001.

[61] Matthew Hausknecht, Prannoy Mupparaju, Sandeep Subramanian, Shivaram Kalyanakrishnan, and Peter Stone. Half field offense: An environment for multiagent learning and ad hoc teamwork. In *AAMAS Adaptive Learning Agents (ALA) Workshop*, 2016.

[62] Milos Hauskrecht, Nicolas Meuleau, Leslie Pack Kaelbling, Thomas Dean, and Craig Boutilier. Hierarchical solution of markov decision processes using macro-actions. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 220–229. Morgan Kaufmann Publishers Inc., 1998.

[63] Ronald A Howard. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26, 1966.

[64] Yuko Ishiwaka, Takamasa Sato, and Yukinori Kakazu. An approach to the pursuit problem on a heterogeneous multiagent system using reinforcement learning. *Robotics and Autonomous Systems*, 43(4):245–256, 2003.

[65] Damian Isla. Handling complexity in the halo 2 ai. In *Game Developers Conference*, volume 12, 2005.

[66] Jonathan Y Ito, David V Pynadath, and Stacy C Marsella. A decision-theoretic approach to evaluating posterior probabilities of mental models. In *AAAI-07 workshop on plan, activity, and intent recognition*, 2007.

[67] Nicholas R Jennings. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *Artificial intelligence*, 75(2):195–240, 1995.

[68] Nick R Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *The knowledge engineering review*, 8(03):223–250, 1993.

[69] W Lewis Johnson. Agents that explain their own actions. *AD-A280 063*, 8:21, 1994.

[70] Nicholas K Jong and Peter Stone. State abstraction discovery from irrelevant state variables. In *IJCAI*, volume 8, pages 752–757, 2005.

[71] Catholijn M Jonker, M Birna Van Riemsdijk, and Bas Vermeulen. Shared mental models. In *Coordination, Organizations, Institutions, and Norms in Agent Systems VI*, pages 132–151. Springer, 2011.

[72] Kshitij Judah, Alan Fern, and Thomas Dietterich. Active imitation learning via state queries. In *Proceedings of the ICML Workshop on Combining Learning Strategies to Reduce Label Cost*. Citeseer, 2011.

[73] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

[74] Thomas Keller and Malte Helmert. Trial-based heuristic tree search for finite horizon mdps. In *ICAPS*, 2013.

[75] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*, pages 282–293. Springer, 2006.

[76] Jelle R Kok, Eter Jan Hoen, Bram Bakker, and Nikos Vlassis. Utile coordination: Learning interdependencies among cooperative agents. In *EEE Symp. on Computational Intelligence and Games, Colchester, Essex*, pages 29–36, 2005.

[77] Sarit Kraus, Katia Sycara, and Amir Evenchik. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence*, 104(1):1–69, 1998.

[78] John Laird and Michael VanLent. Human-level ai's killer application: Interactive computer games. *AI magazine*, 22(2):15, 2001.

[79] Hector J Levesque, Philip R Cohen, and José HT Nunes. On acting together. In *AAAI*, volume 90, pages 94–99, 1990.

[80] Lars Lidén. Artificial stupidity: The art of intentional mistakes. *AI Game Programming Wisdom*, 2:41–48, 2003.

[81] Andreas Lux and Donald Steiner. Understanding cooperation: An agent's perspective. In *ICMAS*, pages 261–268, 1995.

[82] Patrick MacAlpine, Katie Genter, Samuel Barrett, and Peter Stone. The robocup 2013 drop-in player challenges: Experiments in ad hoc teamwork. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 382–387. IEEE, 2014.

[83] Owen Macindoe, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Pomcop: Belief space planning for sidekicks in cooperative games. In *AIIDE*, 2012.

[84] Stacy C Marsella, David V Pynadath, and Stephen J Read. Psychsim: Agent-based modeling of social interactions and influence. In *Proceedings of the international conference on cognitive modeling*, pages 243–248. Citeseer, 2004.

[85] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2):273, 2000.

[86] Kevin McGee and Aswin Thomas Abraham. Real-time team-mate ai in games: A definition, survey, & critique. In *proceedings of the Fifth International Conference on the Foundations of Digital Games*, pages 124–131. ACM, 2010.

[87] Francisco S Melo and Manuela Veloso. Learning of coordination: Exploiting sparse interactions in multiagent systems. In *Proceedings of The 8th*

International Conference on Autonomous Agents and Multiagent Systems-Volume 2, pages 773–780. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[88] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[89] Manisha Mundhe and Sandip Sen. Evaluating concurrent reinforcement learners. In *MultiAgent Systems, 2000. Proceedings. Fourth International Conference on*, pages 421–422. IEEE, 2000.

[90] Kevin P Murphy. A survey of pomdp solution techniques. Technical report, UC Berkeley, 2000.

[91] Ranjit Nair, Milind Tambe, Makoto Yokoo, David Pynadath, and Stacy Marsella. Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *IJCAI*, pages 705–711, 2003.

[92] Ranjit Nair, Pradeep Varakantham, Milind Tambe, and Makoto Yokoo. Networked distributed pomdps: A synthesis of distributed constraint optimization and pomdps. In *AAAI*, volume 5, pages 133–139, 2005.

[93] Hirohide Ushida Yuji Hirayama Hiroshi Nakajima. Emotion model for life-like agent and its evaluation. In *AAAI 98*, page 62. Aaai Pr, 1998.

[94] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsâĂŤi. *Mathematical programming*, 14(1):265–294, 1978.

[95] Brenda Ng, Carol Meyers, Kofi Boakye, and John Nitao. Towards applying interactive pomdps to real-world adversary modeling. In *Innovative Applications in Artificial Intelligence (IAAI)*, pages 1814–1820, 2010.

[96] Truong-Huy Dinh Nguyen, David Hsu, Wee Sun Lee, Tze-Yun Leong, Leslie Pack Kaelbling, Tomas Lozano-Perez, and Andrew Haydn Grant. Capir: Collaborative action planning with intention recognition. In *AIIDE*, 2011.

[97] Muaz Niazi and Amir Hussain. Agent-based computing from multi-agent systems to agent-based models: a visual survey. *Scientometrics*, 89(2):479–499, 2011.

[98] Judith Orasanu. Shared problem models and flight crew performance. *Aviation psychology in practice*, pages 255–285, 1994.

[99] Jeff Orkin. Three states and a plan: the ai of fear. In *Game Developers Conference*, volume 2006, page 4, 2006.

[100] Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, 2005.

[101] Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[102] Ronald Edward Parr. *Hierarchical control and learning for Markov decision processes*. PhD thesis, Citeseer, 1998.

[103] Jim Pitman et al. Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for âĂę, 2002.

[104] Marc JV Ponsen, Geert Gerritsen, and Guillaume Chaslot. Integrating opponent models with monte-carlo tree search in poker. In *Interactive Decision Theory and Game Theory*, 2010.

[105] Ruth Pordes, Don Petravick, Bill Kramer, Doug Olson, Miron Livny, Alain Roy, Paul Avery, Kent Blackburn, Torre Wenaus, Frank Würthwein, et al. The open science grid. In *Journal of Physics: Conference Series*, volume 78, page 012057. IOP Publishing, 2007.

[106] David V Pynadath and Stacy Marsella. Minimal mental models. In *Proceedings Of The National Conference On Artificial Intelligence*, volume 22, page 1038. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[107] David V Pynadath and Stacy C Marsella. Psychsim: Modeling theory of mind with decision-theoretic agents. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, volume 5, pages 1181–1186, 2005.

[108] David V. Pynadath and Milind Tambe. The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of Artificial Intelligence Research*, pages 389–423, 2002.

[109] David V Pynadath and Milind Tambe. 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*, pages 873–880. ACM, 2002.

[110] Miquel Ramírez and Hector Geffner. Goal recognition over pomdps: Inferring the intention of a pomdp agent. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pages 2009–2014. AAAI Press, 2011.

[111] Matthew J Rattigan, Marc Maier, and David Jensen. Exploiting network structure for active inference in collective classification. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, pages 429–434. IEEE, 2007.

[112] Alexander Repenning. Excuse me, i need better ai!: employing collaborative diffusion to make game ai child's play. In *Proceedings of the 2006 ACM SIGGRAPH symposium on Videogames*, pages 169–178. ACM, 2006.

[113] Patrick Riley and Manuela Veloso. Recognizing probabilistic opponent movement models. In *RoboCup 2001: Robot Soccer World Cup V*, pages 453–458. Springer, 2002.

[114] David Robles and Simon M Lucas. A simple tree search method for playing ms. pac-man. In *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pages 249–255. IEEE, 2009.

[115] Maayan Roth, Reid Simmons, and Manuela Veloso. Reasoning about joint beliefs for execution-time communication decisions. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 786–793. ACM, 2005.

[116] Maayan Roth, Reid Simmons, and Manuela Veloso. What to communicate? execution-time decision in multi-agent pomdps. In *Distributed Autonomous Robotic Systems 7*, pages 177–186. Springer, 2006.

[117] Maayan Roth, Reid Simmons, and Manuela Veloso. Exploiting factored representations for decentralized execution in multiagent teams. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 72. ACM, 2007.

[118] William B Rouse, Janis A Cannon-Bowers, and Eduardo Salas. The role of mental models in team performance in complex systems. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(6):1296–1308, 1992.

[119] Deb Roy and Ehud Reiter. Connecting language to the world. *Artificial Intelligence*, 167(1):1–12, 2005.

[120] Trevor Santarra and Arnav Jhala. Communicating intentions for coordination with unknown teammates. In *Proceedings of the 2016 International*

Conference on Autonomous Agents & Multiagent Systems, pages 1423–1424. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[121] Trevor Sarratt and Arnav Jhala. Rapid: A belief convergence strategy for collaborating with inconsistent agents. In *AAAI Workshop on Multiagent Interaction without Prior Coordination*, 2015.

[122] Trevor Sarratt and Arnav Jhala. Tuning belief revision for coordination with inconsistent agents in ad hoc teams (extended abstract). In *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*, 2015.

[123] Trevor Sarratt, David V. Pynadath, and Arnav Jhala. Converging to a player model in monte-carlo tree search. In *Computational Intelligence and Games (CIG), 2014 IEEE Conference on*. IEEE, In press.

[124] Matthias Scheutz. Agents with or without emotions?. In *FLAIRS Conference*, pages 89–93, 2002.

[125] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[126] Igor Sfiligoi, Daniel C Bradley, Burt Holzman, Parag Mhashilkar, Sanjay Padhi, and Frank Wurthwein. The pilot way to grid resources using glideinwms. In *2009 WRI World congress on computer science and information engineering*, volume 2, pages 428–432. IEEE, 2009.

[127] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[128] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

[129] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In *Advances in Neural Information Processing Systems*, pages 2164–2172, 2010.

[130] H Simon. A bounded-rationality model of rational choice. *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting*, 1957.

[131] Luc Steels. Evolving grounded communication for robots. *Trends in cognitive sciences*, 7(7):308–312, 2003.

[132] Reinhard Stolle and Elizabeth Bradley. Multimodal reasoning for automatic model construction. In *AAAI/IAAI*, pages 181–188, 1998.

[133] Peter Stone, Gal A Kaminka, Sarit Kraus, Jeffrey S Rosenschein, and Noa Agmon. Teaching and leading an ad hoc teammate: Collaboration without pre-coordination. *Artificial Intelligence*, 203:35–65, 2013.

[134] Peter Stone, Gal A Kaminka, Sarit Kraus, Jeffrey S Rosenschein, et al. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI*, 2010.

[135] Peter Stone, Gal A Kaminka, and Jeffrey S Rosenschein. Leading a best-response teammate in an ad hoc team. In *Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets*, pages 132–146. Springer, 2010.

[136] Peter Stone and Sarit Kraus. To teach or not to teach?: decision making under uncertainty in ad hoc teams. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 117–124. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[137] Peter Stone and Manuela Veloso. Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. *Artificial Intelligence*, 110(2):241–273, 1999.

[138] Peter Stone and Manuela Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3):345–383, 2000.

[139] Gabriel Synnaeve and Pierre Bessiere. A bayesian model for plan recognition in rts games applied to starcraft. In *Proceedings of the Seventh Artificial Intelligence and Interactive Digital Entertainment International Conference (AIIDE 2011)*, 2011.

[140] Milind Tambe. Recursive agent and agent-group tracking in a real-time dynamic environment. In *ICMAS*, pages 368–375, 1995.

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

[142] Milind Tambe, W Lewis Johnson, Randolph M Jones, Frank Koss, John E Laird, Paul S Rosenbloom, and Karl Schwamb. Intelligent agents for interactive simulation environments. *AI magazine*, 16(1):15, 1995.

[143] Chek Tien Tan and Ho-lun Cheng. An automated model-based adaptive architecture in modern games. In *Proceedings of the Sixth Artificial Intelligence and Interactive Digital Entertainment International Conference (AIIDE 2010)*, 2010.

[144] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

[145] Cagatay Undeger and Faruk Polat. Multi-agent real-time pursuit. *Autonomous Agents and Multi-Agent Systems*, 21(1):69–107, 2010.

[146] Michael Van Lent, William Fisher, and Michael Mancuso. An explainable artificial intelligence system for small-unit tactical behavior. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pages 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[147] Tim Verweij, Martijn Schut, Remco Straatman, and BV Guerrilla. *A hierarchically-layered multiplayer bot system for a first-person shooter*. PhD thesis, Master's thesis, Vrije Universiteit of Amsterdam, 2007.

[148] Rene Vidal, Omid Shakernia, H Jin Kim, David Hyunchul Shim, and Shankar Sastry. Probabilistic pursuit-evasion games: theory, implementation, and experimental evaluation. *IEEE transactions on robotics and automation*, 18(5):662–669, 2002.

[149] Ben George Weber and Michael Mateas. A data mining approach to strategy prediction. In *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pages 140–147. IEEE, 2009.

[150] Andrew Whiten. *Natural theories of mind: Evolution, development and simulation of everyday mindreading*. Basil Blackwell Oxford, 1991.

[151] Michael Wooldridge and Nicholas R Jennings. Towards a theory of cooperative problem solving. In *Distributed Software Agents and Applications*, pages 40–53. Springer, 1996.

[152] John Yen, Xiaocong Fan, Shuang Sun, Timothy Hanratty, and John Dumer. Agents with shared mental models for enhancing team decision makings. *Decision Support Systems*, 41(3):634–653, 2006.

[153] John Yen, Xiaocong Fan, Shuang Sun, Rui Wang, Cong Chen, Kaivan Kamali, and Richard A Volz. Implementing shared mental models for collaborative teamwork. In *the Workshop on Collaboration Agents: Autonomous Agents for Collaborative Environments in the IEEE/WIC Intelligent Agent Technology Conference, Halifax, Canada*, 2003.

[154] Haruhiro Yoshimoto, Kazuki Yoshizoe, Tomoyuki Kaneko, Akihiro Kishimoto, and Kenjiro Taura. Monte carlo go has a way to go. In *AAAI*, volume 6, pages 1070–1075, 2006.

[155] Jiajia Zhang. Building opponent model in imperfect information board games. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 12(3), 2014.