Multiagent Teamwork: Analyzing the Optimality and Complexity of Key Theories and Models

David V. Pynadath and Milind Tambe
Information Sciences Institute and Computer Science Department
University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292
{pynadath,tambe}@isi.edu

ABSTRACT

Despite the significant progress in multiagent teamwork, existing research does not address the *optimality* of its prescriptions nor the *complexity* of the teamwork problem. Thus, we cannot determine whether the assumptions and approximations made by a particular theory gain enough efficiency to justify the losses in overall performance. To provide a tool for evaluating this tradeoff, we present a unified framework, the *COMmunicative Multiagent Team Decision Problem (COM-MTDP)* model, which is general enough to subsume many existing models of multiagent systems. We analyze use the COM-MTDP model to provide a breakdown of the computational complexity of constructing optimal teams under problem domains divided along the dimensions of observability and communication cost. We then exploit the COM-MTDP's ability to encode existing teamwork theories and models to encode two instantiations of joint intentions theory, including STEAM. We then derive a domain-independent criterion for optimal communication and provide a comparative analysis of the two joint intentions instantiations. We have implemented a reusable, domain-independent software package based COM-MTDPs to analyze teamwork coordination strategies, and we demonstrate its use by encoding and evaluating the two joint intentions strategies within an example domain.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*

General Terms

Measurement

1. INTRODUCTION

Multiagent teamwork is critical in a range of domains, including teams of spacecraft, unmanned air vehicles, software agents for logistics planning, and agents for assisting humans. Research into theories of agent teamwork, such as those based on belief-desire-intentions (BDI) frameworks [3, 5, 13], have provided prescriptions for agent coordination. These prescriptions have, in turn, led to practical teamwork models [7, 10, 14, 17] that have succeeded in a range of complex domains. Yet, two key shortcomings limit the scalability of these BDI-based theories and implementations. First, there are no techniques for the quantitative evaluation of the degree of *optimality* of their coordination behavior. While optimal coordination may be impractical in real-world domains, such analysis would aid us in comparison of different existing theories/models and in identifying feasible improvements. One key reason for the difficulty in quantitative evaluation of most existing teamwork theories is that they ignore the various uncertainties and costs in real-world environments. For instance, joint intentions theory [3] prescribes that team members attain mutual beliefs in key circumstances, but it ignores the cost of attaining mutual belief (e.g., via communication). On the other hand, practical systems have addressed costs and uncertainties of real-world environments. For instance, STEAM [14, 15] extends joint-intentions with decision-theoretic communication selectivity. Unfortunately, the very pragmatism of such approaches often necessarily leads to a lack of theoretical rigor, so it remains unanswered whether STEAM's selectivity is the best an agent can do, or whether it is even necessary at all. The second key shortcoming of existing teamwork research is the lack of a characterization of the computational complexity of various aspects of teamwork decisions. Understanding the computational advantages of a practical coordination prescription could potentially justify the use of that prescription as an approximation to optimality in particular domains.

To address these shortcomings, we propose a new framework, the *COMmunicative Multiagent Team Decision Problem (COM-MTDP)*, inspired by work in *economic team theory* [8, 6]. As in that framework, our definition of a team assumes only a common goal (i.e., a joint utility function). Unlike typical teamwork frameworks, we make no other assumptions about the team's behavior (e.g., the teammates form a joint commitment, communicate to attain mutual belief, etc.). We view these more intermediate concepts as the *means* by which agents improve their overall performance, not ends in themselves. For example, while mutual belief has no inherent value, our COM-MTDP model can quantify the improved performance that we would expect from a team that attains mutual belief about important aspects of its execution. While our COM-MTDP model borrows from a theory developed in another field, we make several contributions in applying and extending the original theory, most notably adding explicit models of communication and system dynamics. With these extensions, the COM-MTDP generalizes other recently developed multiagent decision frameworks, such as decentralized POMDPs [1].

This paper demonstrates three new types of teamwork analyses made possible by the COM-MTDP model. First, we analyze the computational complexity of teamwork within problem domains

AAMAS'02, July 15-19, 2002, Bologna, Italy.
Copyright 2002 ACM 1-58113-480-0/02/0007 ...$5.00.

classified along the dimensions of observability and communication cost. Second, the COM-MTDP model provides a powerful tool for analyzing the *optimality* of coordination prescriptions across classes of domains. We encode existing team coordination strategies, based on joint intentions [3] and STEAM [14], within a COM-MTDP for evaluation. We also derive a novel coordination algorithm that outperforms these existing these coordination strategies in optimality, though not in efficiency. The end result is a *well-grounded characterization of the complexity-optimality tradeoff* among various means of team coordination. Third, we can use the COM-MTDP model to empirically analyze a specific domain of interest. We apply our implemented reusable, domain-independent algorithms to empirically evaluate the aforementioned coordination strategies within an example domain represented as a COM-MTDP. We thus characterize the optimality of each strategy as a function of the properties of the underlying domain and, as a result, explain previously unexplained published data.

2. THE COM-MTDP MODEL

2.1 Multiagent Team Decision Problems

Given a team of selfless agents, α, who intend to perform some joint task, we wish to evaluate possible policies of behavior. We represent a *multiagent team decision problem* (MTDP) model as a tuple, $\langle S, \boldsymbol{A}, P, \boldsymbol{\Omega}, \boldsymbol{O}, \boldsymbol{B}, R\rangle$. We have taken the underlying components of this model from the initial team decision model [6], but we have extended them to handle dynamic decisions over time and to more easily represent multiagent domains.

2.1.1 World States: S

S is a set of world states (the team's environment).

2.1.2 Domain-Level Actions: $\boldsymbol{A}$

$\{A_i\}_{i\in\alpha}$ is a set of actions for each agent to perform to change its environment, implicitly defining a set of combined actions, $\boldsymbol{A} \equiv \prod_{i\in\alpha} A_i$ (corresponding team theory's *decision variables*).

Extension to Dynamic Problem: P. The original team decision problem focused on a one-shot, static problem. We extend the original concept so that each component is a time series of random variables. The effects of domain-level actions, as well as any exogenous environmental changes, obey a probabilistic distribution, $P(s_i, \boldsymbol{a}, s_f) = \Pr(S^{t+1} = s_f | S^t = s_i, \boldsymbol{A}^t = \boldsymbol{a})$.

2.1.3 Agent Observations: $\boldsymbol{\Omega}$

$\{\Omega_i\}_{i\in\alpha}$ is a set of observations that each agent, i, can experience of its world, implicitly defining a combined observation, $\boldsymbol{\Omega} \equiv \prod_{i\in\alpha} \Omega_i$. Ω_i may include elements corresponding to indirect evidence of the state (e.g., sensor readings) and actions of other agents. In the original team-theoretic framework, the *information structure* that represented the observation process of the agents was a set of deterministic functions, $O_i : S \to \Omega_i$.

Extension of Allowable Information Structures: $\boldsymbol{O}$. We extend the information structure representation to allow for uncertain observations. We use a general stochastic model, borrowed from the *partially observable Markov decision process* model [11], with a joint observation function: $\boldsymbol{O}(s, \boldsymbol{a}, \boldsymbol{\omega}) = \Pr(\boldsymbol{\Omega}^t = \boldsymbol{\omega} | S^t = s, \boldsymbol{A}^{t-1} = \boldsymbol{a})$. This function models the sensors, representing any errors, noise, etc. In some cases, we can separate this joint distribution into individual observation functions: $\boldsymbol{O} \equiv \prod_{i\in\alpha} O_i$, where $O_i(s, \boldsymbol{a}, \omega) = \Pr(\Omega_i^t = \omega | S^t = s, \boldsymbol{A}^{t-1} = \boldsymbol{a})$. We can distinguish among different classes of information structures:

Collective Partial Observability: This is the general case, where we make no assumptions on the observations.

Collective Observability: There is a unique world state for the *combined* observations of the team: $\forall \boldsymbol{\omega} \in \boldsymbol{\Omega}$, $\exists s \in S$ such that $\forall s' \neq s$, $\Pr(\boldsymbol{\Omega}^t = \boldsymbol{\omega} | S^t = s') = 0$. The set of domains that are collectively observable is a strict subset of the domains that are collectively partially observable.

Individual Observability: There is a unique world state for each individual agent's observations: $\forall \omega \in \Omega_i$, $\exists s \in S$ such that $\forall s' \neq s$, $\Pr(\Omega_i^t = \omega | S^t = s') = 0$. The set of domains that are individually observable is a strict subset of the domains that are collectively observable.

2.1.4 Policy (Strategy) Space

π_{iA} is a domain-level *policy* (or *strategy*, in the original team theory specification) to map an agent's belief state to an action. In the original formalism, the agent's beliefs correspond directly to its observations (i.e., $\pi_{iA} : \Omega_i \to A$).

Extension to Richer Belief State Space: $\boldsymbol{B}$. We generalize the set of possible strategies to capture the more complex mental states of the agents. Each agent, $i \in \alpha$, forms a belief state, $b_i^t \in B_i$, based on its observations seen through time t, where B_i circumscribes the set of possible belief states for the agent. Thus, we define the set of possible domain-level policies as mappings from belief states to actions, $\pi_{iA} : B_i \to A$. We define the set of possible combined belief states over all agents to be $\boldsymbol{B} \equiv \prod_{i\in\alpha} B_i$. The corresponding random variable, $\boldsymbol{b}^t$, represents the agents' combined belief state at time t. We elaborate on different types of belief states and the mapping of observations to belief states (i.e., the *state estimator function*) in Section 2.2.1.

2.1.5 Reward Function: R

A common reward function is central to the notion of teamwork in a MTDP: $R : S \times \boldsymbol{A} \to \mathbb{R}$. This function represents the team's joint preferences over states and the cost of domain-level actions.

2.2 Extension for Explicit Communication: $\boldsymbol{\Sigma}$

We make an explicit separation between domain-level actions (A) and communicative actions. Thus, we extend our initial MTDP model to be a *communicative multiagent team decision problem* (COM-MTDP), that we define as a tuple, $\langle S, \boldsymbol{A}, \boldsymbol{\Sigma}, P, \boldsymbol{\Omega}, \boldsymbol{O}, \boldsymbol{B}, R\rangle$, with a new component, $\boldsymbol{\Sigma}$, and an extended reward function, R

2.2.1 Communication: $\boldsymbol{\Sigma}$

$\{\Sigma_i\}_{i\in\alpha}$ is a set of possible messages for each agent, implicitly defining a set of combined communications, $\boldsymbol{\Sigma} \equiv \prod_{i\in\alpha} \Sigma_i$. An agent, i, may communicate message $x \in \Sigma_i$ to its teammates, who interpret the communication by updating their belief states in response. Thus, the agents now update their belief states at two distinct points within each decision epoch: once upon receiving observation Ω_i^t (producing the *pre-communication* belief state $b_{i\bullet\Sigma}^t$), and again upon receiving the other agents' messages (producing the *post-communication* belief state $b_{i\Sigma\bullet}^t$). The distinction allows us to differentiate between the belief state used by the agents in selecting their communication actions and the more "up-to-date" belief state used in selecting their domain-level actions. We also distinguish between the separate *state-estimator* functions used in each update phase: $b_i^0 = SE_i^0()$, $b_{i\bullet\Sigma}^t = SE_{i\bullet\Sigma}(b_{i\Sigma\bullet}^{t-1}, \Omega_i^t)$, $b_{i\Sigma\bullet}^t = SE_{i\Sigma\bullet}(b_{i\bullet\Sigma}^t, \boldsymbol{\Sigma}^t)$, where $SE_{i\bullet\Sigma} : B_i \times \Omega_i \to B_i$ is the pre-communication state estimator for agent i, and $SE_{i\Sigma\bullet} : B_i \times \boldsymbol{\Sigma} \to B_i$ is the post-communication state estimator for agent i. The initial state estimator, $SE_i^0 : \emptyset \to B_i$, specifies the agent's prior beliefs, before any observations are made. For each of these, we also make the obvious definitions for the corresponding estimators for the combined belief states: $\boldsymbol{SE}_{\bullet\Sigma}$, $\boldsymbol{SE}_{\Sigma\bullet}$, and $\boldsymbol{SE}^0$.

In this paper, we assume that the agents have *perfect recall*, so that the agents recall all of their observations and all communication from other agents. Thus, their belief states can represent the histories of combined observations: $B_i = \Omega_i^* \times \boldsymbol{\Sigma}^*$. The agents realize perfect recall through the following state estimator functions:

$$SE_i^0() = \langle\rangle \tag{1}$$

$$\begin{aligned} & SE_{i\bullet\Sigma}(\langle\langle\Omega_i^0, \boldsymbol{\Sigma}^0\rangle, \ldots, \langle\Omega_i^{t-1}, \boldsymbol{\Sigma}^{t-1}\rangle\rangle, \Omega_i^t) \\ & \quad = \langle\langle\Omega_i^0, \boldsymbol{\Sigma}^0\rangle, \ldots, \langle\Omega_i^{t-1}, \boldsymbol{\Sigma}^{t-1}\rangle, \langle\Omega_i^t, \cdot\rangle\rangle \end{aligned} \tag{2}$$

$$\begin{aligned} & SE_{i\Sigma\bullet}(\langle\langle\Omega_i^0, \boldsymbol{\Sigma}^0\rangle, \ldots, \langle\Omega_i^{t-1}, \boldsymbol{\Sigma}^{t-1}\rangle, \langle\Omega_i^t, \cdot\rangle\rangle, \boldsymbol{\Sigma}^t) \\ & \quad = \langle\langle\Omega_i^0, \boldsymbol{\Sigma}^0\rangle, \ldots, \langle\Omega_i^t, \boldsymbol{\Sigma}^t\rangle\rangle \end{aligned} \tag{3}$$

Note that, although we assume perfect, instantaneous communication here, we could potentially use the post-communication state estimator to model any noise, temporal delays, cognitive burden, etc. present in the communication channel.

We extend our definition of a policy of behavior to include a *coordination policy*, $\pi_{i\Sigma} : B_i \to \Sigma_i$, analogous to Section 2.1.4's domain-level policy. We define the joint policies, $\boldsymbol{\pi}_\Sigma$ and $\boldsymbol{\pi}_A$, as the combined policies across all agents in α.

2.2.2 Extended Reward Function: R

We extend the team's reward function to also represent the cost of communicative acts (e.g., communication channels may have associated cost): $R : S \times \boldsymbol{\Sigma} \times \boldsymbol{A} \to \mathbb{R}$. We assume that the cost of communication and of domain-level actions are independent of each other, so we can decompose the reward function into two components: a communication-level reward, $R_\Sigma : S \times \boldsymbol{\Sigma} \to \mathbb{R}$, and a domain-level reward, $R_A : S \times \boldsymbol{A} \to \mathbb{R}$. The total reward is the sum of the two component values: $R(s, \boldsymbol{\sigma}, \boldsymbol{a}) = R_\Sigma(s, \boldsymbol{\sigma}) + R_A(s, \boldsymbol{a})$. We assume that communication has no inherent benefit and may instead have some cost, so that for all states, $s \in S$, and messages, $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}$, the reward is never positive: $R_\Sigma(s, \boldsymbol{\sigma}) \leq 0$. However, although we assign communication no explicit value, it can have significant, implicit value through its effect on agents belief states, and subsequently on future actions.

As with the observability function, we parameterize the communication costs associated with message transmissions:

General Communication: no assumptions about communication

Free Communication: $R_\Sigma(s, \boldsymbol{\sigma}) = 0$ for any $\boldsymbol{\sigma} \in \boldsymbol{\Sigma}$, and $s \in S$

No Communication: $\boldsymbol{\Sigma} = \emptyset$, i.e., no *explicit* communication

The *free-communication* case appears in the literature, when researchers wish to focus on issues other than communication cost. Although, real-world domains rarely exhibit such ideal conditions, we may be able to model some domains as having approximately free communication to a sufficient degree. In addition, analyzing this extreme case gives us some understanding of the benefit of communication, even if the results do not apply across all domains. We also identify the *no-communication* case because such decision problems have been of interest to researchers as well [4]. Of course, even if $\boldsymbol{\Sigma} = \emptyset$, it is possible that there are domain-level actions in $\boldsymbol{A}$ that have *implicit* communicative value by acting as signals that convey information to the other agents. However, we still label such agent teams as having *no communication* for the purposes of the work here, since many of our results exploit an *explicit* separation between domain- and communication-level actions.

Model	$\boldsymbol{\Sigma}$	$\boldsymbol{O}$
DEC-POMDP	no comm.	collective partial observ.
POIPSG	no comm.	collective partial observ.
MMDP	no comm.	individual observability
Xuan-Lesser	gen. comm.	collective observability

Table 1: Existing models as COM-MTDP subsets.

2.3 Model Illustration

We can view the evolving state as a Markov chain with separate stages for domain-level and coordination-level actions. In other words, each agent team member, $i \in \alpha$ begins in some initial state, S^0, with initial belief states, $b_i^0 = SE_i^0()$. Each agent receives an observation Ω_i^0 drawn according to the probability distribution $\boldsymbol{O}(S^0, \text{null}, \boldsymbol{\Omega}^0)$ (there are no actions yet). Then, each agent updates its belief state, $b_{i\bullet\Sigma}^0 = SE_{i\bullet\Sigma}(b_i^0, \Omega_i^0)$.

Next, each agent $i \in \alpha$ selects a message according to its coordination policy, $\Sigma_i^0 = \pi_{i\Sigma}(b_{i\bullet\Sigma}^0)$, defining a combined communication, $\boldsymbol{\Sigma}^0$. Each agent interprets the communications of all of the others by updating its belief state, $b_{i\Sigma\bullet}^0 = SE_{i\Sigma\bullet}(b_{i\bullet\Sigma}^0, \boldsymbol{\Sigma}^0)$. Each then selects an action according to its domain-level policy, $A_i^0 = \pi_{iA}(b_{i\Sigma\bullet}^0)$, defining a combined action $\boldsymbol{A}^0$. By the central assumption of teamwork, all of the agents receive the same joint reward, $R^0 = R(S^0, \boldsymbol{\Sigma}^0, \boldsymbol{A}^0)$. The world then moves into a new state, S^1, according to the distribution, $P(S^0, \boldsymbol{A}^0)$. Again, each agent i receives an observation Ω_i^1 drawn from Ω_i according to the distribution $\boldsymbol{O}(S^1, \boldsymbol{A}^0, \boldsymbol{\Omega}^1)$, and it updates its belief state, $b_{i\bullet\Sigma}^1 = SE_{i\bullet\Sigma}(b_{i\Sigma\bullet}^0, \Omega_i^1)$. The process continues, with agents choosing communication- and domain-level actions, observing the effects, and updating their beliefs. We can define the *value*, V, of the policies, $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_\Sigma$, as the expected reward received when executing those policies. Over a finite horizon, T, this value is equivalent to the following: $V^T(\boldsymbol{\pi}_A, \boldsymbol{\pi}_\Sigma) = E[\sum_{t=0}^T R^t | \boldsymbol{\pi}_A, \boldsymbol{\pi}_\Sigma]$.

2.4 COM-MTDPs Subsume Existing Models

The COM-MTDP model subsumes many existing multiagent models, as presented in Table 1 (i.e., we can map any instance of these models into a corresponding COM-MTDP). This generality enables us to perform novel analyses of real-world teamwork domains, as demonstrated by Section 4's use of the COM-MTDP model for analyzing the optimality of communication decisions.

With its model of observability and world dynamics, our COM-MTDP model closely parallels the *decentralized partially observable Markov decision process* (DEC-POMDP) [1]. Following our notational conventions, a DEC-POMDP is a tuple, $\langle S, \boldsymbol{A}, P, \boldsymbol{\Omega}, \boldsymbol{O}, R\rangle$. There is no set of possible messages, $\boldsymbol{\Sigma}$, so the DEC-POMDP falls into the class of domains with *no communication*. The DEC-POMDP observational model, $\boldsymbol{O}$, is general enough to capture *collectively partially observable* domains. The *partially observable IPSG* (POIPSG) [9] is very similar to the DEC-POMDP model (i.e., *collectively partially observable* domains, with *no communication*). Like the DEC-POMDP, the *multiagent Markov decision process* (MMDP) [2] has *no communication*. However, the MMDP is a multiagent extension to the completely observable MDP model, so it assumes an environment that is *individually observable*.

The COM-MTDP's separation of communication from other actions is similar to previous work on multiagent decision models [16]. However, while the Xuan-Lesser model generalizes beyond individually observable environments, it supports only a subset of collectively observable environments. In particular, the Xuan-Lesser framework cannot represent agents who receive local observations of a common world state, where the observations of different agents could potentially be interdependent.

3. COM-MTDP COMPLEXITY ANALYSIS

We can use the COM-MTDP model to prove some results about the complexity of constructing optimal agent teams. The problem facing these agents (or the designer of these agents) is how to construct the joint policies, $\boldsymbol{\pi}_\Sigma$ and $\boldsymbol{\pi}_A$, so as to maximize their joint utility, as represented by the expected value, $V^T(\boldsymbol{\pi}_A, \boldsymbol{\pi}_\Sigma)$.

THEOREM 1. *The decision problem of whether there exist policies, $\boldsymbol{\pi}_\Sigma$ and $\boldsymbol{\pi}_A$, for a given COM-MTDP, under* general communication, *that yield a total reward at least K over some finite horizon T is NEXP-complete if $|\alpha| \geq 2$ (i.e., more than one agent).*

Proof: We can reduce a DEC-POMDP to a COM-MTDP with no communication by copying all of the other model features from the given DEC-POMDP. The decision problem for a DEC-POMDP is known to be NEXP-complete [1]. □

In the remainder of this section, we examine the effect of communication on the complexity of constructing optimal team policies. We start by examining the case under the condition of *free communication*, where we would expect the benefit of communication to be the greatest. To begin with, suppose that each agent is capable of communicating its entire observation (i.e., $\Sigma_i \supseteq \Omega_i$). The following theorem tells us that the agents should then exploit this capability and communicate their true observation, as long as they incur no cost in doing so.

THEOREM 2. *Under* free communication, *consider a team of agents using a coordination policy: $\pi_{i\Sigma}(b^t_{i\bullet\Sigma}) \equiv \Omega^t_i$. If the domain-level policy $\boldsymbol{\pi}_A$ maximizes $V^T(\boldsymbol{\pi}_A, \boldsymbol{\pi}_\Sigma)$, then this combined policy is dominant over any other policies. In other words, for all policies, $\boldsymbol{\pi}'_A$ and $\boldsymbol{\pi}'_\Sigma$, $V^T(\boldsymbol{\pi}_A, \boldsymbol{\pi}_\Sigma) \geq V^T(\boldsymbol{\pi}'_A, \boldsymbol{\pi}'_\Sigma)$.*

Proof: Suppose we have some other coordination policy, $\boldsymbol{\pi}'_\Sigma$, that specifies something other than complete communication (e.g., keeping quiet, lying). Suppose that there is some domain-level policy, $\boldsymbol{\pi}'_A$, that allows the team to attain some expected reward, K, when used in combination with $\boldsymbol{\pi}'_\Sigma$. Then, we can construct a domain-level policy, $\boldsymbol{\pi}_A$, such that the team attains the same expected reward, K, when used in conjunction with the complete communication policy, $\boldsymbol{\pi}_\Sigma$, as defined in the statement of Theorem 2.

The coordination policy, $\boldsymbol{\pi}'_\Sigma$, produces a different set of belief states (denoted ${b'}^t_{i\bullet\Sigma}$ and ${b'}^t_{i\Sigma\bullet}$) than those for $\boldsymbol{\pi}_\Sigma$ (denoted $b^t_{i\bullet\Sigma}$ and $b^t_{i\Sigma\bullet}$). In particular, we use state estimator functions, $SE'_{i\bullet\Sigma}$ and $SE'_{i\Sigma\bullet}$ as defined in Equations 2 and 3 to generate ${b'}^t_{i\bullet\Sigma}$ and ${b'}^t_{i\Sigma\bullet}$. Each belief state is a complete history of observation and communication pairs for each agent. On the other hand, under the complete communication of $\boldsymbol{\pi}_\Sigma$, the state estimator functions of Equations 2 and 3 reduce to:

$$SE_{i\bullet\Sigma}(\langle \boldsymbol{\Omega}^0, \ldots, \boldsymbol{\Omega}^{t-1} \rangle, \Omega^t_i) = \langle \boldsymbol{\Omega}^0, \ldots, \boldsymbol{\Omega}^{t-1}, \Omega^t_i \rangle \quad (4)$$

$$\begin{aligned} SE_{i\Sigma\bullet}(\langle \boldsymbol{\Omega}^0, \ldots, \boldsymbol{\Omega}^{t-1}, \Omega^t_i \rangle, \boldsymbol{\Sigma}^t) &= \langle \boldsymbol{\Omega}^0, \ldots, \boldsymbol{\Omega}^{t-1}, \boldsymbol{\Sigma}^t \rangle \\ &= \langle \boldsymbol{\Omega}^0, \ldots, \boldsymbol{\Omega}^{t-1}, \boldsymbol{\Omega}^t \rangle \quad (5) \end{aligned}$$

Thus, $\boldsymbol{\pi}_A$ is defined over a different set of belief states than $\boldsymbol{\pi}'_A$. In order to determine an equivalent $\boldsymbol{\pi}_A$, we must first define a recursive mapping, $\boldsymbol{m}$, that translates the belief states defined by $\boldsymbol{\pi}_\Sigma$ into those defined by $\boldsymbol{\pi}'_\Sigma$:

$$\begin{aligned} m_i(b^t_{i\Sigma\bullet}) =& m_i(\langle b^{t-1}_{i\Sigma\bullet}, \boldsymbol{\Omega}^t \rangle) = m_i(\langle b^{t-1}_{i\Sigma\bullet}, \langle \Omega^t_i, \boldsymbol{\Omega}^t \rangle \rangle) \\ =& \left\langle m_i(b^{t-1}_{i\Sigma\bullet}), \left\langle \Omega^t_i, {\boldsymbol{\Sigma}'}^t \right\rangle \right\rangle = \left\langle m_i(b^{t-1}_{i\Sigma\bullet}), \left\langle \Omega^t_i, \prod_{j\in\alpha} {\Sigma'}^t_j \right\rangle \right\rangle \\ =& \left\langle m_i(b^{t-1}_{i\Sigma\bullet}), \left\langle \Omega^t_i, \prod_{j\in\alpha} \pi'_{j\Sigma}(SE'_{j\bullet\Sigma}(m_j(b^{t-1}_{j\Sigma\bullet}), \Omega^t_j)) \right\rangle \right\rangle \end{aligned} \quad (6)$$

Given this mapping, we then specify: $\pi_{iA}(b^t_{i\Sigma\bullet}) = \pi'_{iA}(m_i(b^t_{i\Sigma\bullet}))$. Executing this domain-level policy, in conjunction with the coordination policy, $\boldsymbol{\pi}_\Sigma$, results in the identical behavior as execution of the alternate policies, ${\boldsymbol{\pi}_A}'$ and ${\boldsymbol{\pi}_\Sigma}'$. Therefore, the team following the policies, $\boldsymbol{\pi}_A$ and $\boldsymbol{\pi}_\Sigma$ will achieve the same expected value of K, as under ${\boldsymbol{\pi}_A}'$ and ${\boldsymbol{\pi}_\Sigma}'$. □

Given this dominance of the complete-communication policy, we can prove that the problem of constructing optimal teams is simpler when communication is free.

THEOREM 3. *The decision problem of determining whether there exist policies, $\boldsymbol{\pi}_\Sigma$ and $\boldsymbol{\pi}_A$, for a given COM-MTDP with* free communication, *that yield a total reward at least K over some finite horizon T is PSPACE-complete.*

Proof: The detailed proof is in the appendix to this paper, available at `http://www.isi.edu/teamcore/COM-MTDP`. From Theorem 2, we need to consider only the complete-communication policy, so that the agents will share their observations and form a coherent joint belief state. Therefore, we can show that the problem is equivalent to the decision problem for a single-agent POMDP. □

THEOREM 4. *The decision problem of determining whether there exist policies, $\boldsymbol{\pi}_\Sigma$ and $\boldsymbol{\pi}_A$, for a given COM-MTDP with* free communication *and* collective observability, *that yield a total reward at least K over some finite horizon T is P-complete.*

Proof: The proof follows that of Theorem 3, but with a reduction to and from the MDP decision problem, rather than the POMDP. □

THEOREM 5. *The decision problem of determining whether there exist policies, $\boldsymbol{\pi}_\Sigma$ and $\boldsymbol{\pi}_A$, for a given COM-MTDP with* individual observability, *that yield a total reward at least K over some finite horizon T (given integers K and T) is P-complete.*

Proof: The proof follows that of Theorem 4, except that we can reduce the problem to and from an MDP regardless of what communication policy the team uses. □

Thus, we have used the COM-MTDP framework to characterize the difficulty of problem domains in agent teamwork along the dimensions of communication cost and observability. Table 2 summarizes our results, which we can use in deciding where to concentrate our energies in attacking teamwork problems. The greatest challenges lie in those domains with either *collective observability* or *partial observability* and with nonzero communication cost. Under *collective observability* and *partial observability*, teamwork without communication is highly intractable, but, with *free communication*, the complexity becomes on par with that of single-agent planning problems. Furthermore, the results from Theorems 3 and 4 hold in any domain where the result from Theorem 2 holds (i.e., when complete communication is the dominant policy). Therefore, while perfectly free communication may be rare, these results show that investment in communication in teamwork can pay off with a significant simplification of optimal teamwork. On the other hand, when the world is *individually observable*, communication makes little difference in performance. It should be noted that even under those conditions where the problem is P-complete, the complexity of optimal teamwork is polynomial in the number of states of the world, which may still be impractically high.

4. EVALUATING COORDINATION

Table 2 shows that providing optimal domain-level and coordination policies for teams is a difficult challenge. Many systems alleviate this difficulty by having domain experts provide the domain-level plans [14]. Then, the problem for the agents reduces to generating the appropriate team coordination, $\boldsymbol{\pi}_\Sigma$, to ensure that they

	Individually Observable	Collectively Observable	Collectively Partially Observ.
No Comm.	P-complete	NEXP-complete	NEXP-complete
Gen. Comm.	P-complete	NEXP-complete	NEXP-complete
Free Comm.	P-complete	P-complete	PSPACE-complete

Table 2: Time complexity of COM-MTDPs.

properly execute the domain-level plans, $\boldsymbol{\pi}_A$. In this section, we use our COM-MTDP framework to analyze joint intentions theory [3], which provides a common basis for many existing approaches to team coordination. Section 4.1 models two particular instantiations of joint intentions taken from the literature [7, 14] as COM-MTDP coordination policies. Section 4.2 analyzes the conditions under which these policies are optimal and provides a third candidate policy that makes communication decisions that are locally optimal within the context of joint intentions. In addition to providing the results for the particular coordination strategies investigated, this section also illustrates a general methodology by which one can use our COM-MTDP framework to encode and evaluate coordination strategies proposed by existing multiagent research.

4.1 Joint Intentions in a COM-MTDP

Joint-intention theory provides a prescriptive framework for multiagent coordination in a team setting. It does not make any claims of optimality in its coordination, but it provides theoretical justifications for its prescriptions, grounded in the attainment of mutual belief among the team members. We can use the COM-MTDP framework to identify the domain properties under which attaining mutual belief is optimal and to quantify precisely how suboptimal the performance will be otherwise.

Joint intentions theory requires that team members jointly commit to a joint persistent goal, G. It also requires that when any team member privately believes that G is achieved (or unachievable or irrelevant), it must then attain mutual belief throughout the team about this achievement (or unachievability or irrelevance). To encode this prescription of joint intentions theory within our COM-MTDP model, we first specify the joint goal, G, as a subset of states, $G \subseteq S$, where the desired goal is achieved (or unachievable or irrelevant). Presumably, such a prescription indicates that joint intentions are not specifically intended for *individually observable* environments, since all of the agents would simultaneously observe that $S^t \in G$, thus attaining mutual belief immediately. Instead, the joint-intention framework aims at domains with some degree of unobservability. In such domains, the agents must signal the other agents, either through communication or some informative domain-level action, to attain mutual belief. However, we can also assume that joint-intention theory does not focus on domains with *free communication*, where Theorem 2 shows that we can simply have the agents communicate everything, all the time, without the need for more complex prescriptions.

The joint-intention framework does not specify a precise communication policy for the attainment of mutual belief. One well-known approach [7] applied joint intentions theory by having the agents communicate the achievement of the joint goal, G, as soon as they believe G to be true. To instantiate the behavior of Jennings' agents within a COM-MTDP, we construct a communication policy, $\boldsymbol{\pi}_\Sigma^J$, that specifies that an agent sends the special message, σ_G, when it first believes that G holds. Following joint intentions' assumption of *sincerity* [12], we require that the agents never select the special σ_G message in any belief state where G is not believed to be true with certainty. We can assume that all of the other agents immediately accept the special message, σ_G, as true, so the team attains mutual belief that G is true immediately upon receiving the message, σ_G. We can construct $\boldsymbol{\pi}_\Sigma^J$ in constant time.

The STEAM algorithm is another instantiation of joint intentions that has had success in several real-world domains [14]. Unlike Jennings' instantiation, the STEAM teamwork model includes decision-theoretic communication selectivity. A domain specification includes two parameters for each joint commitment, G: τ, the probability of miscoordinated termination of G; and C_{mt}, the cost of miscoordinated termination of G. In this context, "miscoordinated termination" means that some agents immediately observe that the team has achieved G while the rest do not. STEAM's domain specification also includes a third parameter, C_c, to represent the cost of communication of a fact (e.g., the achievement of G). Using these parameters, the STEAM algorithm evaluates whether the expected cost of miscoordination outweighs the cost of communication. STEAM expresses this criterion as the following inequality: $\tau \cdot C_{mt} > C_c$. We can define a communication policy, $\boldsymbol{\pi}_\Sigma^S$ based on this criterion: if the inequality holds, then an agent that has observed the achievement of G will send the message, σ_G; otherwise, it will not. We can construct $\boldsymbol{\pi}_\Sigma^S$ in constant time.

4.2 Locally Optimal Policy

Although the STEAM policy is more selective than Jennings', it remains unanswered whether it is *optimally* selective, and researchers continue to struggle with the question of when agents should communicate [17]. The few reports of suboptimal (in particular, excessive) communication in STEAM were characterized as an exceptional circumstance, but it is also possible that STEAM's optimal performance is the exception. We use the COM-MTDP model to derive an analytical characterization of optimal communication here, while Section 5 provides an empirical one.

Both policies, $\boldsymbol{\pi}_\Sigma^J$ and $\boldsymbol{\pi}_\Sigma^S$ consider sending σ_G only when an agent first believes that G has been achieved. Once an agent has the relevant belief, they make different choices, and we consider here what the optimal decision is at this point. The domain is not individually observable, so certain agents may be unaware of the achievement of G. When not sending the σ_G message, these unaware agents may unnecessarily continue performing actions in the pursuit of achieving G. The performance of these extraneous actions could potentially incur costs and lead to a lower utility than one would expect when sending the σ_G message.

The decision to send σ_G or not matters only if the team achieves G *and* one agent comes to know this fact. We define the random variable, K_G, to be the earliest time at which an agent knows this fact. We denote agent A_G as the agent who knows of the achievement at time K_G. If $A_G = i$, for some agent, i, and $K_G = t_0$, then agent i has some pre-communication belief state, $b_{i\bullet\Sigma} = \beta$, that indicates that G has been achieved. To more precisely quantify the difference between agent i sending the σ_G message at time K_G vs. never sending it, we define the following value:

$$\begin{aligned}\Delta^T(t_0, i, \beta) \equiv E\left[\sum_{t=0}^{T-t_0} R^{t_0+t} \middle| \Sigma_i^{t_0} = \sigma_G, K_G = t_0, A_G = i, b_{i\bullet\Sigma}^{t_0} = \beta\right] \\ - E\left[\sum_{t=0}^{T-t_0} R^{t_0+t} \middle| \Sigma_i^{t_0} = \text{null}, K_G = t_0, A_G = i, b_{i\bullet\Sigma}^{t_0} = \beta\right]\end{aligned} \tag{7}$$

We assume that, for all times other than K_G, the agents follow some communication policy, $\boldsymbol{\pi}_\Sigma$, that never specifies σ_G. Thus, Δ^T measures the difference in expected reward that hinges on agent i's specific decision to send or not send σ_G at time t_0. Given this definition, it is locally optimal for agent i to send the special message, σ_G, at time t_0, if and only if $\Delta^T \geq 0$.

We can use the COM-MTDP model to derive an operational expression of $\Delta^T \geq 0$. For simplicity, we define notational shorthand for various sequences and combinations of values. We define a partial sequence of random variables, $X^{<t}$, to be the sequence of random variables for all times before t: X^0, X^1, ... , X^{t-1}. We make similar definitions for the other relational operators. The expression, $(S)^T$, denotes the cross product over states of the world, $\prod_{t=0}^{T} S$, as distinguished from the time-indexed random variable, S^T, which denotes the value of the state at time T. The notation, $s^{\geq t_0}[t]$, specifies the element in slot t within the vector $s^{\geq t_0}$. To simplify the conditioning event on the right-hand side of the two expectations in Equation 7, we define Υ to represent the occurrence of a particular subsequence of world and agent belief states, as follows: $\Pr(\Upsilon\left(\langle t_i, t_f\rangle, s, \boldsymbol{\beta}_{\bullet\Sigma}\right)) \equiv \Pr(S^{\geq t_i, \leq t_f} = s,$ $\boldsymbol{b}_{\bullet\Sigma}{}^{\geq t_i, \leq t_f} = \boldsymbol{\beta}_{\bullet\Sigma} \left| K_G = t_0, A_G = i, b^{t_0}_{i\bullet\Sigma} = \beta\right)$. We define the function, $\beta_{\Sigma\bullet}$, to map a pre-communication belief state into the post-communication belief state that arises from a communication policy: $\beta_{\Sigma\bullet}(\boldsymbol{\beta}_{\bullet\Sigma}, \boldsymbol{\pi}_{\Sigma}) \equiv \boldsymbol{SE}_{\Sigma\bullet}(\boldsymbol{\beta}_{\bullet\Sigma}, \boldsymbol{\pi}_{\Sigma}(\boldsymbol{\beta}_{\bullet\Sigma}))$.

THEOREM 6. *If we assume that, upon achievement of G, no communication other than σ_G is possible, then the condition $\Delta^T(t_0,i,\beta) \geq 0$ holds if and only if:*

$$\sum_{s^{\leq t_0} \in (S)^{t_0}} \sum_{\boldsymbol{\beta}^{\leq t_0}_{\bullet\Sigma} \in \boldsymbol{B}^{t_0}} \Pr(\Upsilon(\langle 0, t_0\rangle, s^{\leq t_0}, \boldsymbol{\beta}^{\leq t_0}_{\bullet\Sigma}))$$

$$\cdot \left(\sum_{s^{\geq t_0} \in (S)^{T-t_0+1}} \sum_{\boldsymbol{\beta}^{\geq t_0}_{\bullet\Sigma} \in \boldsymbol{B}^{T-t_0+1}} \Pr\Big(\Upsilon(\langle t_0, T\rangle, s^{\geq t_0}, \boldsymbol{\beta}^{\geq t_0}_{\bullet\Sigma})\right.$$

$$\Big| \Sigma^{t_0}_i = \sigma_G, \Upsilon(\langle 0, t_0\rangle, s^{\leq t_0}, \boldsymbol{\beta}^{\leq t_0}_{\bullet\Sigma})\Big)$$

$$\cdot \sum_{t=t_0}^{T} R_A\left(s^{\geq t_0}[t], \boldsymbol{\pi}_A\left(\beta_{\Sigma\bullet}\left(\boldsymbol{\beta}_{\bullet\Sigma}{}^{\geq t_0}[t], \boldsymbol{\pi}_{\Sigma}\right)\right)\right)$$

$$- \sum_{s^{\geq t_0} \in (S)^{T-t_0+1}} \sum_{\boldsymbol{\beta}^{\geq t_0}_{\bullet\Sigma} \in \boldsymbol{B}^{T-t_0+1}} \Pr\Big(\Upsilon(\langle t_0, T\rangle, s^{\geq t_0}, \boldsymbol{\beta}^{\geq t_0}_{\bullet\Sigma})$$

$$\Big| \Sigma^{t_0}_i = \textit{null}, \Upsilon(\langle 0, t_0\rangle, s^{\leq t_0}, \boldsymbol{\beta}^{\leq t_0}_{\bullet\Sigma})\Big)$$

$$\left. \cdot \sum_{t=t_0}^{T} R_A\left(s^{\geq t_0}[t], \boldsymbol{\pi}_A\left(\beta_{\Sigma\bullet}\left(\boldsymbol{\beta}_{\bullet\Sigma}{}^{\geq t_0}[t], \boldsymbol{\pi}_{\Sigma}\right)\right)\right) \right)$$

$$\geq - \sum_{s \in G} \sum_{\boldsymbol{\beta} \in \boldsymbol{B}} \Pr\left(\Upsilon(\langle t_0, t_0\rangle, s, \boldsymbol{\beta})\right) R_{\Sigma}(s, \sigma_G) \tag{8}$$

Proof: See `http://www.isi.edu/teamcore/COM-MTDP/`.□

Theorem 6 states, informally, that we prefer sending σ_G whenever the the cost of execution after achieving G outweighs the cost of communication of the fact that G has been achieved. More precisely, the outer summations on the left-hand side of the inequality iterate over all possible past histories of world and belief states, producing a probability distribution over the possible states the team can be in at time t_0. For each such state, the expression inside the parentheses computes the difference in domain-level reward, over all possible *future* sequences of world and belief states, between sending and not sending σ_G. Thus, the left-hand side captures our intuition that, when not communicating, the team will incur a cost if the agents other than i are unaware of G's achievement. The right-hand side of the inequality is a summation of the cost of sending the σ_G message over possible current states and belief states.

	Individually Observable	Collectively Observable	Collectively Partially Obs.
No Comm.	$\Omega(1)$	$\Omega(1)$	$\Omega(1)$
Gen. Comm.	$\Omega(1)$	$O((\lvert S\rvert \cdot \lvert\boldsymbol{\Omega}\rvert)^T)$	$O((\lvert S\rvert \cdot \lvert\boldsymbol{\Omega}\rvert)^T)$
Free Comm.	$\Omega(1)$	$\Omega(1)$	$\Omega(1)$

Table 3: Time complexity of locally optimal decision.

Under *no communication,* we cannot send σ_G. Under *free communication,* the right-hand side is 0, so the inequality is always true, and we know to prefer sending σ_G. Under no assumptions about communication, the determination is more complicated. When the domain is *individually observable*, the left-hand side becomes 0, because *all* of the agents know that G has been achieved (and thus there is no difference in execution when sending σ_G). Therefore, the inequality is always false (unless under *free communication*), and we prefer not sending σ_G. When the environment is not individually observable and communication is available but not free, then, to be locally optimal at time t_0, agent i must evaluate Inequality 8 in its full complexity. Since the inequality sums rewards over all possible sequences of states and observations, the time complexity of the corresponding algorithm is $O((|S| \cdot |\boldsymbol{\Omega}|)^T)$. While this complexity is unacceptable for most real-world problems, it still presents a considerable savings of a factor of $O(2^T)$ over searching the entire policy space for the globally optimal policy, where agent A_G could potentially send σ_G at times other than K_G. Table 3 provides a table of the complexity required to determine the locally optimal policy under the various domain properties.

We can now show that although Theorem 6's algorithm for locally optimal communication provides a significant computational savings over finding the global optimum, it still outperforms existing teamwork models, as exemplified by our $\boldsymbol{\pi}^J_{\Sigma}$ and $\boldsymbol{\pi}^S_{\Sigma}$ policies. First, we can use the criterion of Theorem 6 to evaluate the optimality of the policy, $\boldsymbol{\pi}^J_{\Sigma}$. If $\Delta^T(t_0, i, \beta) \geq 0$ for all possible times t_0, agents i, and belief states β that are consistent with the achievement of the goal G, then the locally optimal policy will *always* specify sending σ_G. In other words, $\boldsymbol{\pi}^J_{\Sigma}$ will be identical to the locally optimal policy. However, if the inequality of Theorem 6 is *ever* false, then $\boldsymbol{\pi}^J_{\Sigma}$ is not even locally, let alone globally, optimal.

Second, we can also use Theorem 6 to evaluate STEAM by viewing STEAM's inequality, $\tau \cdot C_{mt} > C_c$, as a crude approximation of Inequality 8. In fact, there is a clear correspondence between the terms in the two inequalities. The left-hand side of Inequality 8 computes an exact expected cost of miscoordination. However, unlike STEAM's monolithic τ parameter, the optimal criterion evaluates a complete probability distribution over all possible states of miscoordination by considering all possible past sequences consistent with the agent's current beliefs. Likewise, unlike STEAM's monolithic C_{mt} parameter, the optimal criterion looks ahead over all possible future sequences of states to determine the true expected cost of miscoordination. Furthermore, we can view STEAM's parameter, C_c, as an approximation of the communication cost computed by the right-hand side of Inequality 8. Again, STEAM uses a single parameter, while the optimal criterion computes an expected cost over all possible states of the world. On the other hand, the optimal criterion derived with the COM-MTDP model provides a justification for the overall structure behind STEAM's approximate criterion. Furthermore, STEAM's emphasis on on-line computation makes the computational complexity of Inequality 8 (as presented in Table 3) unacceptable), so the approximation error may be acceptable given the gains in efficiency. For a specific domain, we can use empirical evaluation (as demonstrated in the next section) to quantify the error and efficiency to precisely judge this tradeoff.

5. EMPIRICAL POLICY EVALUATION

In addition to providing these analytical results over general classes of problem domains, the COM-MTDP framework also supports the analysis of *specific* domains. Given a particular problem domain, we can construct an optimal coordination policy or, if the complexity of computing an optimal policy is prohibitive, we can instead evaluate and compare candidate approximate policies. To provide a reusable tool for such evaluations, we have implemented the COM-MTDP model as a Python class with domain-independent methods for the evaluation of arbitrary policies and for the generation of both locally optimal policies using Theorem 6 and globally optimal policies through brute-force search of the policy space.

This section presents results of a COM-MTDP analysis of an example domain involving agent-piloted helicopters, where we isolate a single decision, but vary the cost of communication and degree of observability to generate a space of distinct domains with different implications for the agents' performance. By evaluating communication policies over various configurations of this particular testbed domain, we demonstrate a methodology by which one can use the COM-MTDP framework to model *any* problem domain and to evaluate candidate coordination policies for it.

Consider two helicopters that must fly across enemy territory to their destination. The first, piloted by agent T, is a transport vehicle, and the second, piloted by agent E, is an escort vehicle. An enemy radar unit is along their path, but neither agent knows the location a priori. E can destroy the radar unit upon encountering it, but T cannot. Given its superior firepower, E does not worry about detection; therefore, it will fly at its normal speed and altitude. T, on the other hand, must escape radar detection by traveling at a very low altitude (*nap-of-the-earth* flight) and at a lower speed than at its typical, higher altitude. Once E has destroyed the radar, it is then safe for T to fly at its normal altitude and speed.

The two agents form a top-level joint commitment, G_D, to reach their destination. There is no reason for the agents to communicate the achievement of this goal. However, in the service of their top-level goal, G_D, the two agents also adopt a joint commitment, G_R, of destroying the radar unit. We consider here the problem facing E with respect to communicating the achievement of goal, G_R. If E communicates the achievement of G_R, then T knows that it is safe to fly at its normal altitude (thus reaching the destination sooner). If E does *not* communicate the achievement of G_R, there is still some chance that T will observe the event anyway. If T does not observe the achievement of G_R, then it must fly nap-of-the-earth the whole distance, and the team receives a lower reward because of the later arrival. Therefore, E must weigh the increase in expected reward against the cost of communication.

In the COM-MTDP model of this scenario, the world state is the position of T, E, and the enemy radar. The enemy is at a randomly selected position somewhere in between the agents' initial position and their destination. T has no possible communication actions, but it can choose between two domain-level actions: flying nap-of-the-earth and flying at its normal speed and altitude. E has two domain-level actions: flying at its normal speed and altitude, or destroying the radar. E also has the option of communicating the special message, σ_{G_R}, indicating that the radar has been destroyed.

If E arrives at the radar, then it observes its presence with certainty and destroys it, achieving G_R. The likelihood of T's observing the radar's destruction is a function of its distance from the radar. We can vary this function's *observability* parameter within the range $[0, 1]$ to generate distinct domain configurations (0 means that T will never observe the radar's destruction; 1 means T will always observe it). If the observability is 1, then they achieve mutual belief of the achievement of G_R as soon as it occurs. However, for any observability less than 1, there is a chance that the agents will not achieve mutual belief simply by common observation. The helicopters receive a fixed reward for each time step spent at their destination. Thus, for a fixed time horizon, the earlier the helicopters reach there, the greater the team's reward. Since flying nap-of-the-earth is slower than normal speed, T will switch to its normal flying as soon as it either observes that G_R has been achieved or E sends the message, σ_{G_R}. Sending the message is not free, so we impose a variable communication cost, also within the range $[0, 1]$.

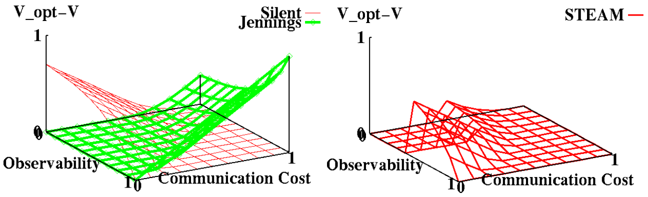

Figure 1: Suboptimality of approximate policies.

We constructed COM-MTDP models of this scenario for each combination of observability and communication cost within the range $[0, 1]$ at 0.1 increments. For each combination, we applied the Jennings and STEAM policies, as well as a completely silent policy. For this domain, the policy, $\boldsymbol{\pi}_{\Sigma}^{J}$, dictates that E always communicate σ_{G_R} upon destroying the radar. For STEAM, we fix the cost of miscoordination, C_{mt}, but vary the τ and C_c parameters with the observability and communication cost parameters, respectively. Following the published STEAM algorithm [14], E sends message σ_{G_R} if and only if STEAM's inequality $\tau \cdot C_{mt} > C_c$, holds. We also constructed locally and globally optimal policies.

Figure 1 plots how much utility the team can expect to lose by following the Jennings, silent, and STEAM policies instead of the locally optimal communication policy (thus, higher values mean *worse* performance). We can immediately see that the Jennings and silent policies are significantly suboptimal for many possible domain configurations. For example, not surprisingly, the surface for the policy, $\boldsymbol{\pi}_{\Sigma}^{J}$, peaks (i.e., it does most poorly) when the communication cost is high and when the observability is high, while the silent policy does poorly under exactly the opposite conditions.

Figure 1 shows the expected value lost by following the STEAM policy. We can view STEAM as trying to intelligently interpolate between the Jennings and silent policies based on the particular domain properties. In fact, we see two thresholds, one along each dimension, at which STEAM switches between following the Jennings and silent policies, and its suboptimality is highest at these thresholds. Thus, its performance generally follows the better of those two fixed policies, so its maximum suboptimality (0.587) is significantly lower than that of the silent (0.700) and Jennings' (1.000) policies. Furthermore, STEAM outperforms the two policies on average, across the space of domain configurations, as evidenced by its mean suboptimality of 0.063, which is less than half of the silent policy's mean of 0.160 and the Jennings' policy's mean of 0.161. Thus, we have been able to quantify the savings provided by STEAM over less selective policies within this example domain.

However, within a given domain configuration, STEAM must either always or never communicate, and this inflexibility leads to significant suboptimality. We see STEAM's limitations more clearly in Figure 2, which plots the expected number of messages sent using STEAM vs. the locally optimal policy, at an observability of 0.3. STEAM's expected number of messages is either 0 or 1, so STEAM can make at most two (instantaneous) transitions between them: one threshold value each along the observability and communication cost dimensions. Figure 2 shows that the optimal policy can be more flexible than STEAM by specifying communication contingent on E's beliefs beyond simply the achievement of G_R. For example, even if the communication cost is high, it is still worth sending message σ_{G_R} in states where T is still very far from the destination. Thus, the surface for the optimal policy, makes a more gradual transition from always communicating to never communicating. We can thus view STEAM's surface as a crude approximation to the optimal surface, subject to STEAM's fewer degrees of freedom.

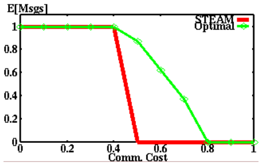

Figure 2: Expected number of messages.

We can also use Figure 2 to identify the domain conditions under which joint-intentions theory's prescription of attaining mutual belief is or is not optimal. In particular, for any domain where the observability is less than 1, the agents will not attain mutual belief without communication. In Figure 2, there are *many* domain configurations where the locally optimal policy is expected to send fewer than 1 σ_{G_R} message. Each of these configurations represents a domain where the locally optimal policy will not attain mutual belief in at least one case. Therefore, attaining mutual belief is suboptimal in those configurations!

These experiments illustrate that STEAM, despite its decision-theoretic communication selectivity, may communicate suboptimally under a significant class of domain configurations. Previous work on STEAM-based, real-world, agent-team implementations informally noted suboptimality in an isolated configuration within a more realistic helicopter transport domain [14]. Unfortunately, this previous work treated that suboptimality (where the agents communicated more than necessary) as an isolated aberration, so there was no investigation of the degree of such suboptimality, nor of the conditions under which such suboptimality may occur in practice. We re-created these conditions within the experimental testbed of this section by increasing the STEAM parameter, C_{mt}, representing the cost of miscoordination. The resulting experiments (graphs omitted for space) illustrated that the observed suboptimality was not an isolated phenomenon, but, in fact, that STEAM has a general propensity towards extraneous communication in situations involving low observability (i.e., low likelihood of mutual belief) and high communication costs. This result matches the situation where the "aberration" occurred in the more realistic domain.

The locally optimal policy is itself suboptimal with respect to the globally optimal policy. Under domain configurations with high observability, the globally optimal policy has the escort wait an additional time step after destroying the radar and then communicate only if the transport continues flying nap-of-the-earth. This leads to a slight advantage in expected utility over the locally optimal policy, with a mean difference of 0.011, standard deviation of 0.027, and maximum of 0.120 (full graph omitted for space). On the other hand, our domain-independent code never requires more than 5 seconds to compute the locally optimal policy in this testbed, while generating the globally optimal policy required more than 150 *minutes*. Thus, through Theorem 6, we have used the COM-MTDP model to construct a communication policy that, for this testbed domain, performs almost optimally and outperforms existing teamwork theories, with a substantial computational savings over finding the globally optimal policy.

6. CONCLUSION

The COM-MTDP model is a novel framework that complements existing teamwork research by providing the previously lacking capability to analyze the optimality and complexity of team decisions. While grounded within economic team theory, the COM-MTDP's extensions to include communication and dynamism allow it to subsume many existing multiagent models. We were able to exploit the COM-MTDP's ability to represent broad classes of multiagent team domains to derive complexity results for optimal agent teamwork under arbitrary problem domains. We also used the model to identify domain properties that can simplify that complexity.

The COM-MTDP framework provides a general methodology for analysis across both general domain subclasses and specific domain instantiations. As demonstrated in Section 4, we can express important existing teamwork theories within a COM-MTDP framework and derive broadly applicable theoretical results about their optimality. Section 5 demonstrates our methodology for the analysis of a specific domain. By encoding a teamwork problem as a COM-MTDP, we can use the leverage of our general-purpose software tools (available at `www.isi.edu/teamcore/COM-MTDP`) to evaluate the optimality of coordination based on potentially any other existing teamwork theory, as demonstrated in this paper using two leading teamwork theories: joint intentions and STEAM. In combining both theory and practice, we can use the theoretical results derived using the COM-MTDP as the basis for new algorithms to extend our software tools, just as we did in translating Theorem 6 from Section 4 into an implemented algorithm for locally optimal communication in Section 5. We expect that the COM-MTDP framework, the theorems and complexity results, and the reusable software will form a basis for further analysis of teamwork, both by ourselves and others in the field.

Acknowledgments. This research was supported by DARPA award No F30602-98-2-0108, under the Control of Agent Based Systems program, and managed by AFRL/Rome Research Site. We would like to thank Ashish Goel, Daniel Marcu, and Paul Rosenbloom for valuable discussion and feedback.

7. REFERENCES

[1] D.S. Bernstein, S. Zilberstein, & N. Immerman. The complexity of decentralized control of MDPs. *UAI*, 32–37, 2000.

[2] C. Boutilier. Planning, learning & coordination in multiagent decision processes. *TARK*, 195–210, 1996.

[3] P.R. Cohen & H.J. Levesque. Teamwork. *Nous*, 25(4):487–512, 1991.

[4] D. Goldberg & M.J. Matarić. Interference as a tool for designing & evaluating multi-robot controllers. *AAAI*, 637–642, 1997.

[5] B. Grosz & S. Kraus. Collab. plans for complex group actions. *AIJ*, 86:269–358, 1996.

[6] Y.-C. Ho. Team decision theory & information structures. *Proc. of the IEEE*, 68(6):644–654, 1980.

[7] N. Jennings. Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. *AIJ*, 75:195–240, 1995.

[8] J. Marschak & R. Radner. *The Econ. Theory of Teams*. Yale, 1971.

[9] L. Peshkin, K.-E. Kim, N. Meuleau, & L.P. Kaelbling. Learning to cooperate via policy search. *UAI*, 489–496, 2000.

[10] C. Rich & C. Sidner. COLLAGEN: When agents collaborate with people. *Conf. on Auton. Agents*, 284–291, 1997.

[11] R.D. Smallwood & E.J. Sondik. The optimal control of POMDPs over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

[12] I.A. Smith & P.R. Cohen. Toward a semantics for an ACL based on speech-acts. *AAAI*, 24–31, 1996.

[13] E. Sonenberg, G. Tidhar, E. Werner, D. Kinny, M. Ljungberg, & A. Rao. Planned team activity. Tech. Rep. 26, Austral. AI Inst., 1994.

[14] M. Tambe. Towards flexible teamwork. *JAIR*, 7:83–124, 1997.

[15] M. Tambe & W. Zhang. Towards flexible teamwork in persistent teams. *ICMAS*, 277–284, 1998.

[16] P. Xuan, V. Lesser, & S. Zilberstein. Communication decisions in multi-agent cooperation. *Conf. on Auton. Agents*, 616–623, 2001.

[17] J. Yen, J. Yin, T.R. Ioerger, M.S. Miller, D. Xu, & R.A. Volz. CAST: Collab. agents for simulating teamwork. *IJCAI*, 1135–1142, 2001.