POLITECNICO DI TORINO
Repository ISTITUZIONALE

Asking for Information by Evaluating the Communication and Coordination Trade-Off in Multi-Agent POMDPs

Original
Asking for Information by Evaluating the Communication and Coordination Trade-Off in Multi-Agent POMDPs / Trombetta, E.I., Capello, E., Shell, D.A., Rossi, F.. - In: IEEE ROBOTICS AND AUTOMATION LETTERS. - ISSN 2377-3766. - 11:8(2026), pp. 9144-9150. [10.1109/LRA.2026.3703246]

Availability:
This version is available at: 11583/3013129 since: 2026-07-15T09:53:39Z

Publisher:
Institute of Electrical and Electronics Engineers

Published
DOI:10.1109/LRA.2026.3703246

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

26 November 2026

Asking for Information by Evaluating the Communication and Coordination Trade-Off in Multi-Agent POMDPs

Enza I. Trombetta, Elisa Capello, *Member, IEEE*, Dylan A. Shell, and Federico Rossi

***Abstract*—This letter concerns decentralized multi-agent systems in which individual agents operate under partial observability and uncertainty, but with the option of invoking inter-agent communication. When information exchange is infeasible or prohibitively expensive —owing to bandwidth limits, energy constraints, or the demands of clandestine operation— one is prevented from continually maintaining a joint belief over the underlying system state. In such cases, the decision of when to request information from others is key. In revisiting the prior work addressing this decision head-on, we see conservative strategies: agents are constrained to act solely on information that is accessible to all, with useful local observations being ignored whenever communication is too costly. Such an approach, though apposite to high-risk settings when consistent decision-making must be guaranteed, can degrade overall performance significantly. We explore how agents may integrate their local information, even when it has not been globally shared, and then communicate only when prudent—we term this an *ask-type* paradigm. This re-formulation, in which the agents are not compelled to synchronize, leads to a novel method that enables the decentralized execution of a centralized policy by reasoning, at execution time, about whether the cost of exchanging information will be offset by the information's value. We empirically evaluate our approach across several benchmarks and a case study, assessing in each case its effectiveness in trading between tolerating some loss of coordination and reducing communication costs.**

***Index Terms*—Planning under uncertainty, multi-robot systems, planning, scheduling and coordination.**

I. Introduction

DECISION-MAKING under uncertainty for multiple agents is a focus area of research that continues to grow owing to its increasing relevance to applications such as target tracking, surveillance, sensor networks, cooperative robotics, intelligent transportation, task allocation, collision-free navigation, and cooperative exploration, among others [1]. *Cooperative decision-making* refers to the challenge of deriving a set of local behaviors for a group of agents operating within an environment to maximize a global reward. In decentralized systems, individual decision-makers lack access to global information and must instead rely on their local histories of observations and actions while reasoning about the potential behaviors of other agents. This problem is formalized by the Decentralized Partially Observable Markov Decision Process (Dec-POMDP) framework [2], which generalizes the Partially Observable Markov Decision Process (POMDP) [3] to a multi-agent context. In practice, the strict decentralization assumption can oftentimes be relaxed, allowing agents to exchange information whenever doing so is beneficial. During mission execution, inter-agent communication serves as a coordination mechanism through which agents synchronize their beliefs, mitigate uncertainty, and improve performance, potentially approaching the effectiveness of a fully centralized planning while still maintaining decentralized execution. When communication is free and broadcasted, the problem reduces to a large centralized POMDP, known as a Multi-agent Partially Observable Markov Decision Process (MPOMDP), where the fully centralized policy is executed through continuous information exchange. Unfortunately, continuous communication can be infeasible or undesirable due to bandwidth limitations, energy constraints, or the risk of detection in hostile or contested environments.

A. Related Works

Existing solution methods, both model-based and learning-based, seek to optimize action-selection and communication either during offline planning or online execution. In this paper, we investigate model-based approaches grounded in decision theory, and we focus on planning strategies that explicitly account for communication to (i) enable decentralized execution of a centralized policy, or (ii) enhance performance during decentralized policy execution.

Deciding *when* to communicate is crucial to avoid bottlenecks, wasted computational resources, and redundant information. Existing approaches include treating communication as an available action in the framework of Dec-POMDPs, which can be solved offline [4], [5] or online [6]; agent-centric interactive POMDP formulations, where agents recursively reason about other agents' beliefs [7], [8], [9]; game-theoretic approaches [10]; and learning approaches, in which agents learn when to communicate through interaction with the environment [11].

An important early piece of work, that of [12], approaches the problem of decentralized coordination by deliberating, during execution, over whether it is beneficial to share some information with other agents (they do assume that free communication is available at planning time). Agents maintain a joint belief that is consistent with information that has been shared with all. When an agent learns something new —e.g., disclosed via a recent local observation— it determines, depending on whether the communication cost can be justified, if it is worth everyone knowing it or not. When the answer is not to communicate, the agent with the new information acts as if itself were unaware (see Fig. 1). This approach, wherein the onus to be consistent with one's peers overrides one's own local knowledge, is *conservative*; it is an appropriate choice

Received 29 December 2025; accepted 3 May 2026. Date of publication 12 June 2026; date of current version 22 June 2026. This article was recommended for publication by Associate Editor B. Yan and Editor C.-B. Yan upon evaluation of the reviewers' comments. This work was supported by ONR under Award N00014-22-1-2476. *(Corresponding author: Enza I. Trombetta.)*

Enza I. Trombetta and Elisa Capello are with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, 10129 Torino, Italy (e-mail: enza.trombetta@polito.it; elisa.capello@polito.it).

Dylan A. Shell is with Texas A&M University, College Station, TX 77840 USA (e-mail: dshell@cs.tamu.edu).

Federico Rossi is with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109 USA (e-mail: federico.rossi@jpl.nasa.gov).

Digital Object Identifier 10.1109/LRA.2026.3703246

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

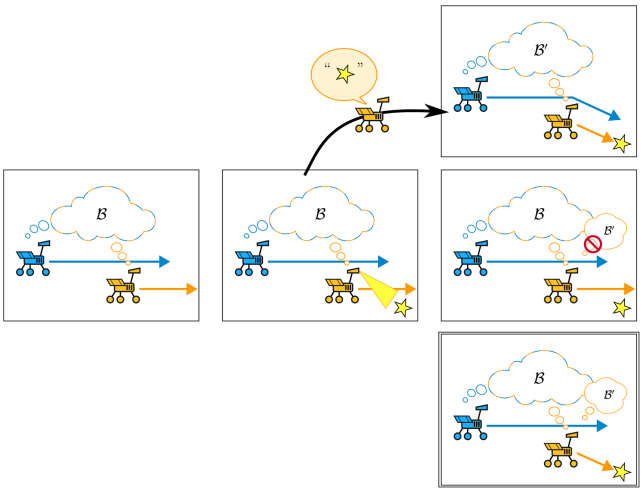


Fig. 1. Two agents coordinating, starting with identical beliefs. The lower agent makes an observation, and determines whether communicating to update the shared belief and subsequent actions (top branch) yields sufficient expected benefit relative to the communication cost. If the cost outweighs the benefit, the agent ignores the observation and continues deriving the joint action from the previously shared belief (middle branch). We examine (double box at right) when the agent uses its local belief, even though this may cause deviations from what is known collectively.

in high-risk circumstances. But can we obtain more efficient behavior by allowing agents to use local information? The present paper establishes that there are task domains in which agents updating their behavior on the basis of local information, even at the potential risk of misalignment, can be beneficial so long as the agents reason about the impact of such misalignment. Indeed, several standard benchmark problems are seen to have this property.

B. Contribution and Organization

Our contributions are summarized as follows:

- Explicitly addressing whether it is beneficial to share information, we formulate an *ask*-type communication paradigm, contrary to [12] that focuses on when to *tell*. For us, each agent preserves an individual perspective during execution, integrating local information, and only requests information from the team when necessary.
- Taking advantage of this shift in perspective, we introduce a novel algorithm that enables decentralized execution of a centralized policy by reasoning, at each time step, about communication. By adopting a *broadcast query strategy*, our method focuses on optimizing *when* to communicate, enabling agents to reduce uncertainty and improve local action selection while minimizing communication overhead.
- Finally, we evaluate the proposed algorithm on benchmarks from the literature and on a case study, showing its effectiveness in managing the trade-off between coordination loss and communication cost. While primarily theoretical, these results highlight its potential applicability to robotic systems and other real-world multi-agent domains.

The remainder of this paper is organized as follows: Section II presents the problem formulation; Section III describes the proposed approach; Section IV reports and discusses experimental results; and Section V concludes the paper.

II. Problem Formulation

Throughout this work, we consider a decentralized multi-agent system composed of rational, cooperative agents operating in a stochastic, partially observable environment. Agents possess local information and may establish a *joint belief* about the global state of the system through communication, which is assumed to be always available, instantaneous, and reliable, but costly. In the following sections, we formally introduce the planning problem, describe belief space generation and propagation, and analyze the proposed strategy to solve the planning problem and mitigate its complexity.

A. System Model

An MPOMDP (or, equivalently, a Dec-POMDP) is defined as a tuple:

$$\langle I, S, A, T, O, \mathcal{O}, R, b_0, h, \gamma \rangle,$$

where $I = \{1, \dots, n\}$ is the set of n agents, S is a finite set of states, summarizing all the relevant features of the agents and their environment, $A = \times_{i \in I} A_i$ is the finite set of joint actions and $O = \times_{i \in I} O_i$ the finite set of joint observations. (The $\times$ operator denotes a Cartesian product.) A joint action for the current time step t is denoted as $a_t = \langle a_{1,t}, \dots, a_{n,t} \rangle$. Similarly, $o_t = \langle o_{1,t}, \dots, o_{n,t} \rangle$ denotes a joint observation for the current time step t. Every time step, agent i only observes its own component o_i.[1] Let $\langle o_i, o_{-i} \rangle$ denote the set of joint observations with o_i the local observation and $o_{-i} \in O_{-i} = \times_{j \in I,\ j \neq i} O_j$ observations of the other agents except i.[2] The transition probability function and the observation probability function are denoted as T and $\mathcal{O}$, respectively. In particular, $T(s'|s, a)$ denotes the probability that taking joint action a in state s results in a transition to state s'. $\mathcal{O}(o|s', a)$ denotes the probability of obtaining joint observation o after taking joint action a and reaching state s'. Transition and observation probabilities are assumed to be stationary. The reward function $R(s, a)$ defines the reward received when the system executes joint action a in state s. The initial state distribution or *belief* at time $t = 0$ is denoted by $b_0 \in \Delta(S)$, with $\Delta(S)$ the set of probability distributions over the set of states. All agents share the initial belief and have complete knowledge of the system model, including the transition and observation functions. Finally, h and $\gamma \in [0, 1)$ are the time horizon and the discount factor, respectively.

The distinctive feature of an MPOMDP is the ability of each agent to compute a *joint belief*,

$$b_t = \Pr(s_t | b_0, a_0, o_1, a_1, \dots, a_{t-1}, o_t),$$

defined as the probability distribution over states given the histories of joint actions and observations [2].

An action-observation history (AOH) for agent i at time step t, denoted $\vec{\theta}_{i,t}$, is the sequence of actions taken by and observations received by agent i,

$$\vec{\theta}_{i,t} = (a_{i,0}, o_{i,1}, \dots, a_{i,t-1}, o_{i,t}).$$

We introduce the notation $\vec{\theta}_{i,t_k \to t}$ to denote the AOH from time t_k to t, with $t_k \in [0, t-1]$, representing the sequence $(a_{i,t_k}, o_{i,t_k+1}, \dots, a_{i,t-1}, o_{i,t})$. Agent i's set of possible AOHs at time t is $\vec{\Theta}_{i,t}$. A joint action-observation history $\vec{\theta}_t$ specifies the AOH for all agents at time t, while the set of possible joint AOHs for all agents is denoted as $\vec{\Theta}_t$. Similar definitions are used for the joint observation history (OH).

The joint belief b_t serves as a *sufficient statistic* for the system, enabling agents to select actions based on the current state of knowledge. Its computation can be done incrementally using Bayes' rule in the same way as in a POMDP, i.e., through a Bayesian belief update function $\tau(b_{t-1}, a_{t-1}, o_t)$ that computes b_t from b_{t-1}, a_{t-1}, and o_t.

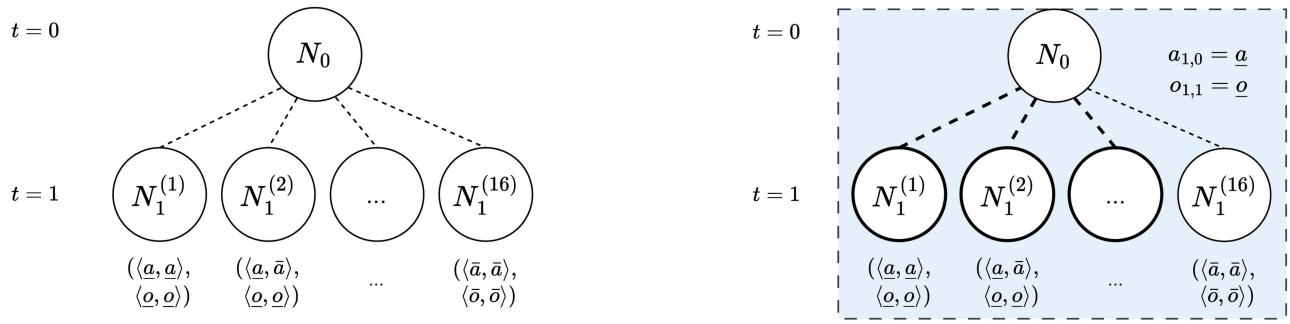
Solving an MPOMDP involves defining a plan, or *policy*, for each agent that maximizes the expected cumulative reward:

$$E\left[\sum_{t=0}^{h-1} R(s_t, a_t) | b_0, \pi\right],$$

also referred to as the *value* V^π of a joint policy π, where the expectation is over states and observations.

[1] For conciseness, the time index is omitted.

[2] With the notation $-i$, we denote the set of all agents except agent i.



(a) System overview

(b) Local overview, Agent 1

Fig. 2. Representation of the branching structure generated to account for all possible joint AOH. A common initial root N_0 represents the initial state of a system composed of two agents with local action and observation spaces $A_i = (\underline{a}, \bar{a})$, and $O_i = (\underline{o}, \bar{o})$, respectively, $i = 1, 2$. On the left, all possible action-observation combinations (Fig. 2(a)). On the right, a local view of Agent 1 isolating the branches consistent with its own local AOH (Fig. 2(b)).

The policy $\pi = (\delta_0, \ldots, \delta_{h-1})$ maps joint beliefs to joint actions at each time step t, $\delta_t : \Delta(S) \to A \ \forall t$, and can be decomposed into individual policies $\pi_i = (\delta_{i,0}, \ldots, \delta_{i,h-1})$, with $\delta_{i,t} : \Delta(S) \to A_i \ \forall t$ for all agents. In Section II-C, we discuss methods for identifying the optimal policy, highlighting the dual challenge of optimizing policy computation and execution whilst simultaneously minimizing costly communication.

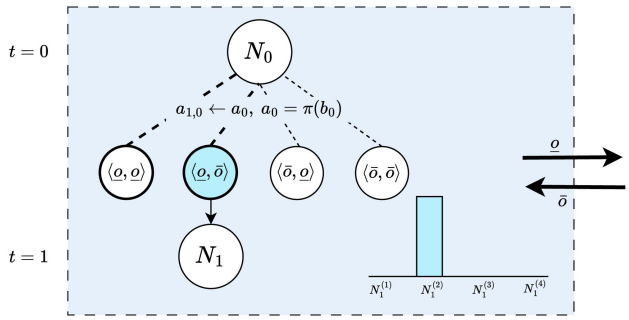
B. Belief Space Generation and Propagation

To identify the optimal action, each agent must reconstruct a joint belief over the system state by exchanging local information with other agents at every time step. If this exchange is not possible, agents cannot fully reconstruct their teammates' AOHs and, as a result, cannot determine a unique joint belief. They can only identify possible joint AOHs —and therefore possible joint beliefs— which can be visualized as branches of a tree starting from a common root (Fig. 2), where each path represents a specific joint history, its resulting joint belief, and its likelihood of occurrence.

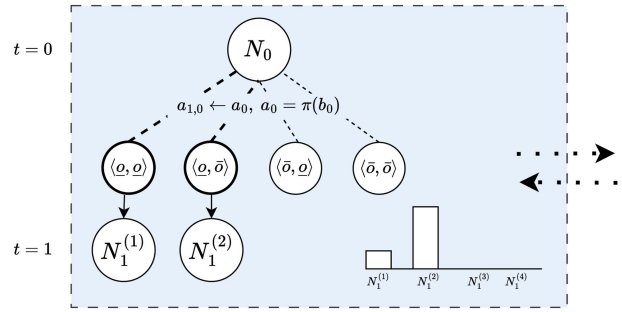
For each agent i, let $\mathcal{B}_t^i$ denote the set of possible joint beliefs compatible with its local information. This set is constructed by agent i based on its own AOH and on reasoning about the possible AOHs of the other agents. At time $t = 0$, all $\mathcal{B}_0^i$ are identical and consist solely of the initial belief b_0. From the joint belief b_0, agent i can derive an optimal joint action a_0 from the centralized policy π and execute its corresponding individual action $a_{i,0}$. The system then transitions to a new state, and each agent receives a local observation $o_{i,1}$.

If the agents decide to communicate (Fig. 3(a)), they can reconstruct the joint AOH and update the joint belief accordingly. All sets $\mathcal{B}_1^i$ are restricted to the current joint belief $b_1 = \tau(b_0, a_0, o_1)$ about the current joint AOH $\vec{\theta}_1 = (a_0, o_1)$, whose probability of occurrence is $\Pr(\vec{\theta}_1) = 1$. If agents do not communicate (Fig. 3(b)), the exact joint belief cannot be constructed, but a set of possible joint beliefs can be inferred. Agent i might naturally do one of two things. First, (i) act blindly and enumerate all possible joint OHs from the root and all possible joint AOHs otherwise–the root constitutes a special case as a_0 is coordinated. Or (ii) incorporate its local information and infer the possible OHs of the remaining agents—in subsequent steps, their AOHs. The first option ensures that all agents reason from the same information–coordination is guaranteed, but valuable local information can be wasted. The second option leverages local information to improve decision-making—even at the potential cost of temporary misalignment— but it allows agents to make more informed choices. Motivated by the question of whether permitting agents to rely on local information can yield more efficient behavior, we focus on the second option.

Agent i estimates a set of possible joint AOHs $\vec{\Theta}_1^i$, and consequently a set of possible joint beliefs, $\mathcal{B}_1^i$, along with their probabilities of occurrence $P_1^i = \{\Pr(\vec{\theta}) \mid \vec{\theta} \in \vec{\Theta}_1^i\}$, by applying $\tau(\cdot)$ to all joint AOHs



(a) With communication



(b) Without communication

Fig. 3. Belief Space Propagation for the first time step, with and without communication, for a team of two agents. The perspective is that of Agent 1, executing local action $a_{1,0}$ and receiving local observation $o_{1,1} = \underline{o}$.

consistent with agent i's local AOH. Algorithm 1 describes, starting at time $t-1$ from the sets $\vec{\Theta}_{t-1}^i$, $\mathcal{B}_{t-1}^i$ and P_{t-1}^i, how to advance them to time t for a given joint action a_{t-1} and a local observation $o_{i,t}$. These sets serve as the basis for agent i to reason about action selection at the next step.

At each step, given a joint action, the belief space grows according to the number of joint observations compatible with agent i's local observation, which in turn depends on the size of the individual agents' observation spaces and the number of agents. In the absence of communication, this space expands exponentially as the number of steps increases. This rapid expansion poses a significant challenge for planning, ultimately rendering the direct enumeration of AOHs infeasible in practice. Moreover, when agents rely solely on local information, their belief sets gradually diverge, increasing the risk of poor system

Algorithm 1: Belief Space Propagation for Agent i.i.

function EXPAND($\vec{\Theta}^i_{t-1}, \mathcal{B}^i_{t-1}, P^i_{t-1}, a_{t-1}, o_{i,t}$)
Initialize $\vec{\Theta}^i_t = \emptyset$, $\mathcal{B}^i_t = \emptyset$, $P^i_t = \emptyset$
for $\vec{\theta}_{t-1} \in \vec{\Theta}^i_{t-1}$ **do**
  Select b_{t-1} from $\mathcal{B}^i_{t-1}$, $\Pr(\vec{\theta}_{t-1})$ from P^i_{t-1}
  for $o_{-i} \in O_{-i} = \times_{j\in I,\ j\neq i} O_j$ **do**
    $\vec{\theta}_t = (\vec{\theta}_{t-1}, a_{t-1}, \langle o_{i,t}, o_{-i}\rangle)$
    $b_t = \tau(b_{t-1}, a_{t-1}, \langle o_{i,t}, o_{-i}\rangle)$
    $\Pr(\vec{\theta}_t) = \Pr(\vec{\theta}_{t-1}) \Pr(\langle o_{i,t}, o_{-i}\rangle | a_{t-1}, b_{t-1})$
    $\vec{\Theta}^i_t \leftarrow \vec{\theta}_t$, $\mathcal{B}^i_t \leftarrow b_t$, $P^i_t \leftarrow \Pr(\vec{\theta}_t)$
  end for
end for
return $\vec{\Theta}^i_t,\ \mathcal{B}^i_t,\ P^i_t$

performance owing to inconsistencies in the agents' views. The effect of communication is precisely to collapse these sets and compute a single joint belief for the system, restoring a common view for all agents.

C. Solving the Planning Problem

Solving Dec-POMDPs with general communication yields a truly decentralized policy, which each agent can execute independently using only local information. However, this approach is computationally demanding, posing significant challenges for both exact and approximate solution methods [13]. By assuming free and instantaneous communication, the multi-agent planning problem is reduced to an MPOMDP, whose complexity is PSPACE [3], compared to the NEXP complexity[3] that arises when trying to solve a Dec-POMDP with general communication [14]. This reduction allows the use of state-of-the-art POMDP solvers, such as [15], [16], [17], to efficiently compute a *centralized policy* for the team. While the computational complexity of offline planning is significantly reduced, the main drawback is the need for continuous and potentially costly information exchange during execution, which may be impractical in real-world scenarios with communication constraints.

The present work's central tenet is to leverage inter-agent communication to execute the centralized policy in a decentralized manner. While the approach does not yield the optimal solution to the Dec-POMDP, it can provide high-quality solutions to a large family of problems where coordination between agents is only required at critical junctures—providing a method to identify such junctures, and communicate only then. We provide a strategy in which each agent integrates local information, reasons about other agents' histories to construct a joint belief set compatible with local information, and applies a decision rule for selecting locally optimal actions. For such a strategy, the crux is then determining whether acquiring additional information from other agents could increase the overall system reward, avoiding penalties that result from miscoordination.

III. Reasoning About Communication

Assume a centralized policy π for the underlying MPOMDP has been generated during the offline planning phase. In this section, we present our approach by addressing three key questions: (i) how agent i selects a local optimal joint action given the set of estimated joint beliefs and the policy π, (ii) how this selection relates to the optimal joint actions identified by the other agents, and (iii) under which conditions communication is unnecessary at a given time step.

A. Identification of the Optimal Action: Local Perspective

Agent i determines the optimal joint action a^i_t by reasoning over its local information and, using it, then derives its local component $a^i_{i,t}$ for execution at time step t. The selection is performed given the set of estimated joint beliefs and the centralized policy. A common approximation consists of selecting a representative belief $\hat{b}$, e.g., the mean belief or the most likely belief of the set, and computing the action as $a^i_t = \pi(\hat{b})$. Alternatively, the agent may adopt a risk-averse strategy by selecting the action that maximizes the worst-case value over the belief set, providing safety guarantees in high-risk settings. However, the representative-belief approximation may result in suboptimal action selection when the belief pool exhibits high diversity, whereas the max–min strategy can induce excessive conservatism and reduced performance in nominal conditions.

To overcome these limitations, we adopt the Q-POMDP heuristic to determine the joint action that maximizes the expected reward over the distribution of possible joint beliefs, assuming that agents communicate their AOHs at the next time step and at all future time steps. It is based on the calculation of the *one-step lookahead value function* Q that defines the expected reward for taking a joint action a in the current time step, at belief b_t, and acting according to the centralized policy in the future:

$$Q(b_t, a) = \sum_{s\in S} R(s,a) b_t(s) + \gamma \sum_{o\in O} \Pr(o|a, b_t) V^\pi(\tau(b_t, a, o)).$$

The local optimal joint action without communication for agent i, $a^i_{t,NC}$, is selected by maximizing the expected reward over the distribution of all possible posterior joint beliefs:

$$a^i_{t,NC} = \arg\max_{a\in A} \sum_{\vec{\theta}_t \in \vec{\Theta}^i_t} \Pr(\vec{\theta}_t) Q(b_t(\vec{\theta}_t), a), \tag{1}$$

with $b_t(\vec{\theta}_t) \in \mathcal{B}^i_t$ joint belief associated with the joint AOH $\vec{\theta}_t$. The expected value for the system evaluated by agent i, given by the execution of $a^i_{t,NC}$, is:

$$V^i_{t,NC} = \sum_{\vec{\theta}_t \in \vec{\Theta}^i_t} \Pr(\vec{\theta}_t) Q(b_t(\vec{\theta}_t), a^i_{t,NC}). \tag{2}$$

If the estimated joint belief sets coincide across all agents, the induced action selection is identical and yields a unique optimal joint action. However, due to the dependence on local information, agents may construct different belief sets, which can lead to discrepancies in the selected joint actions.

B. Identification of the Optimal Action: Global Perspective

We are interested in analyzing how agent i evaluates its own joint action relative to those selected by the other agents, under the assumption that each acts locally based on its own belief set while following a shared centralized policy, decision rule, and common knowledge. Specifically, we reason about how agent i models the behavior of its teammates and incorporates this reasoning into its local decision-making.

A possible approach is the robust strategy, in which agent i assumes no knowledge about the other agents' AOH and expands its estimated joint belief set considering all possible actions and observations of the other agents. In this case, the Q-POMDP heuristic selects an action that is "good" across all possible behaviors of the other agent, avoiding overconfidence but potentially limiting performance due to its conservatism. Conversely, agent i can assume that the other agents will behave according to its identified local optimal joint action, effectively narrowing the set of possible AOHs. This can provide computational advantages and higher rewards when the belief pools align, but it also introduces a significant risk of failure if the assumptions about the other agents' behavior are incorrect. We refer to this strategy as optimistic, reflecting the agent's expectation that the other agents will act in accordance with its own reasoning. Between the fully uncertain view and the optimistic view, intermediate approaches can leverage an expected-utility decision rule over the local belief pool while considering a restricted subset of plausible actions for the other agents. However, they may require evaluating multiple candidate actions, thereby increasing the computational complexity and extending the processing time.

[3] Complexity analysis is valid for finite-horizon problems.

In this paper, we examine the strengths and limitations of adopting the optimistic approach, motivated by the observation that in domains characterized by predictable, aligned, or weakly coupled agent behaviors, arising from high agent independence or strong alignment of objectives, precise step-by-step knowledge of other agents' actions is often unnecessary, provided agents reason about the potential impact of misalignment to act effectively.

C. When to Communicate: Proposed Approach

The decision of when to communicate is determined by computing the *Value of Information* (VOI), which quantifies the expected benefit of receiving information from other agents relative to the associated communication cost. Information-theoretic approaches measure this benefit as the expected information gain from other agents' observations, normalized by communication cost; regret-based heuristics evaluate the expected loss in cumulative reward incurred when an agent acts without communication; risk- or safety-oriented methods aim to minimize the probability of low-reward or catastrophic outcomes under worst-case assumptions.

In this paper, we define the VOI as the expected gain in reward obtained by accessing other agents' observations at the current and future time steps. Specifically, each agent i evaluates the expected return of executing its local optimal action without communication, $V^i_{t,NC}$, and estimates the expected return $V^i_{t,C}$ if it had access to all agents' observations using the Q-POMDP heuristic:

$$V^i_{t,C} = \sum_{\vec{\theta}_t \in \vec{\Theta}^i_t} \Pr(\vec{\theta}_t) \max_{a \in A} \Big[Q(b_t(\vec{\theta}_t), a) \Big]. \tag{3}$$

The VOI is then calculated as the gain in the expected return, and communication is triggered only when this gain exceeds the cost:

$$|V^i_{t,C} - V^i_{t,NC}| \geq C_{\text{comm}}.$$

This approach ensures that communication decisions remain consistent with the same expected-return criterion used for action selection and can be computed directly from the existing belief representation, without requiring additional computations or manually tuned thresholds.

Algorithm 2 describes the execution phase from the perspective of agent i, highlighting the processes of belief space propagation, VOI evaluation, and optimal action selection. Before execution, agent i initializes its set of possible joint beliefs with the initial state distribution, which represents common knowledge across all agents (line 1). It further initializes variables to record its local AOH, and the information received at the last communication event, i.e., (i) the most recent joint belief reconstructed through communication, (ii) the corresponding joint AOH, and (iii) the time step at which communication occurred (line 2). At line 3, agent i initializes two binary flags: the first, τ_{ask}, is true if agent i requires information from other agents, and the second, τ_{tell}, is true if other agents request information from agent i. At $t = 1$, each agent executes its local action, obtained from the joint action specified by the centralized policy π. This action is *coordinated*, in the sense that it is consistent across all agents. Upon receiving a local observation, agent i updates its local AOH and constructs the set of possible joint histories, joint beliefs, and their associated probabilities with the propagation function (lines 7–8). It then computes the action and the value that would be obtained by acting solely on local information, comparing it with the expected value attainable if the observations of other agents were known. Communication is considered beneficial if the expected improvement exceeds the communication cost. If this condition holds, agent i requests the local AOHs of the other agents since the last communication step, reconstructs the current joint belief using Algorithm 3, and selects the joint action prescribed by the centralized policy. Otherwise, the action without communication is selected. Moreover, it checks whether communication has been requested by other agents (line 21). If so, it transmits its local AOH starting from the step corresponding to the most recent communication request issued by a specific agent.

Algorithm 2: Communication Planning For Agent i.i.

1: Initialize $\vec{\Theta}^i_0 = \emptyset$, $\mathcal{B}^i_0 = b_0$, $P^i_0 = 1.0$, local AOH $\vec{\theta}_{i,0} = \emptyset$
2: Initialize $\mathcal{C} = \langle b_{t_c} = b_0, \vec{\theta}_{t_c} = \emptyset, t_c = 0 \rangle$
3: $\tau_{\text{ask}} = \text{false}$, $\tau_{\text{tell}} = \text{false}$
4: $a_0 \leftarrow \pi(b_0)$
5: **for** $t = 1, 2, \ldots, h$ **do**
6: Execute local action $a_{i,t-1}$, receive local observation $o_{i,t}$
7: $\vec{\theta}_{i,t} = (\vec{\theta}_{i,t-1}, a_{i,t-1}, o_{i,t})$
8: $\vec{\Theta}^i_t, \mathcal{B}^i_t, P^i_t \leftarrow \text{EXPAND}(\vec{\Theta}^i_{t-1}, \mathcal{B}^i_{t-1}, P^i_{t-1}, a_{t-1}, o_{i,t})$
9: Calculate $a^i_{t,NC}$ from (1) and $V^i_{t,NC}$ from (2)
10: Calculate $V^i_{t,C}$ from (3)
11: **if** $|V^i_{t,C} - V^i_{t,NC}| \geq C_{\text{comm}}$ **then**
12: $\tau_{\text{ask}} = \text{true}$
13: Ask for other agents' AOH $\vec{\theta}_{-i,t_c \to t}$
14: $b_t, \vec{\theta}_t \leftarrow \text{RESTORE}(\mathcal{C}, \vec{\theta}_{i,t}, \vec{\theta}_{-i,t_c \to t}, t)$
15: $b_{t_c} \leftarrow b_t$, $\vec{\theta}_{t_c} = \vec{\theta}_t$, $t_c \leftarrow t$
16: Calculate $a_t \leftarrow \pi(b_t)$
17: $\vec{\Theta}^i_t = \vec{\theta}_t$, $\mathcal{B}^i_t = b_t$, $P^i_t = 1.0$
18: **else**
19: $a_t \leftarrow a^i_{t,NC}$
20: **end if**
21: **if** Outstanding requests for comms **then**
22: $\tau_{\text{tell}} = \text{true}$
23: Send local AOH since last request at time t_r, $\vec{\theta}_{i,t_r \to t}$
24: **end if**
25: $\tau_{\text{ask}} = \text{false}$, $\tau_{\text{tell}} = \text{false}$
26: **end for**

Algorithm 3: Joint Belief Recovery At t from t_c, Agent i.

1: **function** RESTORE($\mathcal{C}, \vec{\theta}_{i,t}, \vec{\theta}_{-i,t}, t$)
2: $b_{t_c}, \vec{\theta}_{t_c}, t_c \leftarrow \mathcal{C}$
3: **for** $k = t_c$ **to** $t - 1$ **do**
4: Extract a_i and o_i from $\vec{\theta}_{i,k}$
5: Extract a_{-i} and o_{-i} from $\vec{\theta}_{-i,k}$
6: $\vec{\theta}_{k+1} = (\vec{\theta}_k, \langle a_i, a_{-i} \rangle, \langle o_i, o_{-i} \rangle)$
7: $b_{k+1} = \tau(b_k, \langle a_i, a_{-i} \rangle, \langle o_i, o_{-i} \rangle)$
8: **end for**
9: **return** $b_t, \vec{\theta}_t$

In many practical domains, exact VOI computation can be prohibitively expensive, even under a myopic formulation. When the space of possible joint beliefs is large, sampling-based Monte Carlo methods can be employed to improve computational tractability and flexibility, at the cost of approximation error and potential suboptimality.

IV. Experiments

We evaluate the proposed algorithm on several Dec-POMDP benchmarks, comparing it with reference baselines that capture two extreme behaviors: a fully coordinated strategy and a more conservative one. Specifically, we assess the planning solutions produced by the ALWAYS-COMM strategy (**ACOM**), corresponding to the MPOMDP solution, an algorithm with the same logic as the ACE-PJB-Comm algorithm [18] (**APJC**), and our proposed approach (**OURS**). A centralized policy is computed offline using the SARSOP algorithm [15] and subsequently executed at runtime by the three different strategies. Performance is evaluated over N Monte Carlo (MC) simulation runs, according to the following metrics:

- average cumulative reward r over the MC runs (with 95% confidence intervals);
- average frequency of communication f_c (%);

TABLE I
BENCHMARK PROBLEM SPECIFICATIONS

Problem	n	$\|S\|$	$\|A\|$	$\|O\|$	max_steps	N
Multi-Agent Tiger	2	2	9	4	6	100
Meeting in a Grid	2	16	25	16	5	15
Mars Rovers	2	256	36	64	15	15
Warehouse	2	100	25	256	15	50

TABLE II
PERFORMANCE FOR THE MULTI-AGENT TIGER PROBLEM

Method	r	f_c (%)	t (s)	$\dim(\vec{\Theta}')$	$f_{a_{\neq}}$ (%)
ACOM$_{0.0}$	$\mathbf{7.460 \pm 4.673}$	100.0	0.000	1.0	0.0
APJC$_{0.0}$	6.560 ± 4.223	44.5	0.029	131.8	0.0
OURS$_{0.0}$	$\mathbf{7.460 \pm 4.673}$	100.0	0.013	1.0	0.0
ACOM$_{0.8}$	$\mathbf{6.084 \pm 4.696}$	100.0	0.000	1.0	0.0
APJC$_{0.8}$	5.084 ± 3.287	25.27	0.031	491.08	0.0
OURS$_{0.8}$	$\mathbf{6.084 \pm 4.696}$	100.0	0.000	1.0	0.0
ACOM$_{2.0}$	4.020 ± 4.740	100.0	0.000	1.0	0.0
APJC$_{2.0}$	$\mathbf{4.040 \pm 3.256}$	25.27	0.031	491.08	0.0
OURS$_{2.0}$	-3.850 ± 6.546	47.5	0.000	3.3	15.97
ACOM$_{10.0}$	$\mathbf{-9.740 \pm 5.297}$	100.0	0.000	1.0	0.0
APJC$_{10.0}$	-12.000 ± 0.000	0.00	0.135	3172.0	0.0
OURS$_{10.0}$	-12.890 ± 8.661	0.00	0.000	9.16	19.50

- average computational time t per run, averaged over MC runs, required to decide whether to communicate and what action to take next;
- maximum dimension of the set of estimated joint AOHs over which each algorithm has reasoned to decide the optimal joint action (i.e., the maximum dimension of the joint belief set $\mathcal{B}_t^i$ before a decision on communication has been made), averaged across the MC runs $\dim(\vec{\Theta}')$;
- average frequency of disagreement about the joint action $f_{a_{\neq}}$ (%) (i.e., the frequency with which the optimal joint action locally selected by an agent differs from that of the others), a measure of the *action inconsistency*.

Computational times were measured using a demonstrative, non-optimized implementation, with utilities from the *POMDPs.jl* library [19]. All experiments were run on a PC with an Intel(R) Core(TM) Ultra 7 CPU with 32.0 GB RAM. They are reported solely to illustrate relative differences in runtime across strategies and are not intended as performance benchmarks. For compactness, we denote each method–cost pair as M-$_{C_{\text{comm}}}$, where M- indicates the method and C_{comm} the communication cost. Benchmark problems include *Multi-Agent Tiger Problem* [20], *Meeting in a Grid* 2×2 [21], and *Mars Rovers* [22], with their parameters in Table I , where the notation $|X|$ is used to denote the number of elements in the space X. The benchmarks are chosen to provide a clear, controlled, and reproducible setting for experimentation. To demonstrate the effectiveness of the approach in a more realistic setting, we also introduce the *Warehouse Problem*, a variant of the *Meeting in a Grid* problem, where two agents navigate a 3×4 warehouse grid cluttered with obstacles to reach a common rendezvous location. Each agent perceives only local boundary information (i.e., whether movement is blocked in the four cardinal directions), resulting in partial observability and ambiguity in localization. The obstacle layout induces bottlenecks and alternative routes, requiring agents to coordinate both navigation and timing despite the absence of direct observation of each other, mirroring real-world warehouse robotics, where autonomous agents must coordinate under uncertainty and limited communication to meet at a designated handoff point.

To prevent the estimated joint belief sets from growing excessively, probabilities P_t^i below a threshold of 10^{-4} are pruned from the belief set. The pruning threshold is set empirically to balance efficiency and retention of relevant low-probability hypotheses, but performance is sensitive and may need domain-specific tuning.

TABLE III
PERFORMANCE FOR THE MEETING IN A GRID 2×2 PROBLEM

Method	r	f_c (%)	t (s)	$\dim(\vec{\Theta}')$	$f_{a_{\neq}}$ (%)
ACOM$_{0.0}$	$\mathbf{3.333 \pm 0.563}$	100.0	0.004	1.0	0.0
APJC$_{0.0}$	$\mathbf{3.333 \pm 0.563}$	50.00	0.235	2.2	0.0
OURS$_{0.0}$	$\mathbf{3.333 \pm 0.563}$	56.67	0.025	1.0	0.0
ACOM$_{0.2}$	2.533 ± 0.563	100.0	0.000	1.0	0.0
APJC$_{0.2}$	2.480 ± 0.687	48.33	0.411	196.87	0.0
OURS$_{0.2}$	$\mathbf{2.680 \pm 0.783}$	56.67	0.002	1.8	6.67
ACOM$_{0.4}$	1.733 ± 0.563	100.0	0.001	1.0	0.0
APJC$_{0.4}$	1.440 ± 0.626	18.33	2.137	905.4	0.0
OURS$_{0.4}$	$\mathbf{2.400 \pm 0.953}$	0.00	0.034	278.0	80.0
ACOM$_{0.8}$	0.133 ± 0.563	100.0	0.001	1.0	0.0
APJC$_{0.8}$	1.733 ± 0.619	0.00	3.227	1396.0	0.0
OURS$_{0.8}$	$\mathbf{2.400 \pm 0.953}$	0.00	0.037	278.0	80.0

TABLE IV
PERFORMANCE FOR THE MARS ROVERS PROBLEM

Method	r	f_c (%)	t (s)	$\dim(\vec{\Theta}')$	$f_{a_{\neq}}$ (%)
ACOM$_{0.0}$	$\mathbf{11.787 \pm 1.739}$	100.0	0.005	1.0	0.0
APJC$_{0.0}$	$\mathbf{11.787 \pm 1.739}$	32.22	1.982	2.2	0.0
OURS$_{0.0}$	$\mathbf{11.787 \pm 1.739}$	28.9	0.020	1.0	0.0
ACOM$_{0.2}$	10.907 ± 1.550	100.0	0.002	1.0	0.0
APJC$_{0.2}$	11.573 ± 1.748	32.22	4.382	2.2	0.0
OURS$_{0.2}$	$\mathbf{11.587 \pm 1.739}$	28.9	0.038	1.0	0.0
ACOM$_{2.5}$	0.787 ± 0.675	100.0	0.005	1.0	0.0
APJC$_{2.5}$	2.027 ± 2.396	8.3	7.505	32.0	0.0
OURS$_{2.5}$	$\mathbf{7.440 \pm 3.465}$	18.9	2.322	4.0	17.77
ACOM$_{10.0}$	-32.213 ± 7.798	100.0	0.005	1.0	0.0
APJC$_{10.0}$	-0.973 ± 1.260	0.00	11.353	44.0	0.0
OURS$_{10.0}$	$\mathbf{-0.613 \pm 3.299}$	10.0	2.201	4.0	41.33

TABLE V
PERFORMANCE FOR THE WAREHOUSE PROBLEM

Method	r	f_c (%)	t (s)	$\dim(\vec{\Theta}')$	$f_{a_{\neq}}$ (%)
ACOM$_{0.0}$	$\mathbf{93.360 \pm 8.653}$	100.0	0.002	1.0	0.0
APJC$_{0.0}$	93.160 ± 8.675	17.7	0.413	56.7	0.0
OURS$_{0.0}$	93.160 ± 8.675	36.0	0.012	2.0	0.3
ACOM$_{0.4}$	87.760 ± 8.653	100.0	0.001	1.0	0.0
APJC$_{0.4}$	89.706 ± 9.875	19.0	0.221	87.2	0.0
OURS$_{0.4}$	$\mathbf{91.448 \pm 8.892}$	30.6	0.007	7.0	2.4
ACOM$_{1.0}$	79.360 ± 8.653	100.0	0.001	1.0	0.0
APJC$_{1.0}$	85.980 ± 10.943	20.4	0.412	214.6	0.0
OURS$_{1.0}$	$\mathbf{88.920 \pm 9.318}$	28.9	0.014	46.1	6.4
ACOM$_{5.0}$	23.360 ± 8.653	100.0	0.003	1.0	0.0
APJC$_{5.0}$	22.340 ± 12.342	11.0	1.659	850.0	0.0
OURS$_{5.0}$	$\mathbf{86.760 \pm 8.091}$	8.9	0.346	377.3	31.7

A. Results

Simulation results are presented in Tables II, III, IV and V. In domains with tightly coupled or safety-critical dynamics, such as the *Multi-Agent Tiger Problem*, coordination between agents is crucial, as significant penalties are incurred if they fail to select the same action at each decision point. Our algorithm often opts to communicate because the risk associated with miscoordination is substantial. When communication becomes prohibitively expensive, whereas a conservative approach, such as **APJC** , continues to listen indefinitely, in our algorithm, the agents may choose to open a door as a consequence of the integration of local observations combined with an optimistic estimate of the other agent's behavior. This can lead to either high or very low rewards depending on the simulation outcome, resulting in a substantially higher variance in the returns.

In domains with weakly coupled rewards or low penalties for misaligned actions, such as the *Meeting in a Grid*, *Mars Rovers*, and *Warehouse* problems, our algorithm effectively identifies high-reward actions without communicating at every step. As shown in the last

column of Tables III, IV and V, even without enforced coordination as in **APJC** algorithm, agents either (i) select the same joint action and act in coordination solely by reasoning about the behavior and observations of the other agents, reflecting the validity of the optimistic assumption, or (ii) judge that acting without coordination is not risky, and the high rewards confirm the effectiveness of their choice. Beyond rewards, we also analyze computational aspects, which highlight a key advantage of our algorithm, specifically looking at the t and $\dim(\vec{\Theta}')$ metrics. As the joint belief set grows, computational demands rise quickly, limiting real-time execution. Integrating locally available information and leveraging the optimistic approach to joint action selection substantially reduces computational effort for belief set management, enabling faster Q-POMDP evaluation and reduced memory requirements. For very large trees, scalability is still an issue. Prior work has attempted to address this challenge by constraining communication frequency—e.g., requiring agents to exchange information at least once every K steps [5]— or by pruning and clustering histories to control memory growth [10], [23]. These techniques are fully compatible with our framework and can be incorporated to further improve scalability.

Beyond minimal belief tree expansion, another key feature of the proposed approach is its agent-centric design, which enables each agent to reason locally in a single pass about both action selection and communication, without requiring negotiation or multi-round protocols. In addition, our recovery mechanism ensures that when communication is triggered, the exact joint belief is reconstructed from the true joint AOH, removing any accumulated errors from violations of the assumptions and allowing the agent to restart from a correct belief. A potential limitation arises when communication is prohibitively expensive: in such cases, agents may judge communication as unnecessary and rely excessively on optimistic assumptions, which can be problematic in high-risk environments where a more conservative strategy would be preferable. Moreover, the myopic nature of the decision-making process may also affect the communication trigger, as agents evaluate communication mainly based on short-term benefits. As a consequence, potentially advantageous communication opportunities that would only reveal their benefit over a longer planning horizon may be missed. Considering longer-term effects could instead allow agents to identify more suitable moments for communication, where an immediate cost leads to better coordination and lower uncertainty in the future. Investigating communication policies that explicitly account for longer-term effects, therefore, represents an important direction for future work.

V. Conclusion

This paper has returned to an early and influential idea in decentralized coordination, namely, deliberating about whether communication actions are valuable. We offer a shift in perspective from the impactful work of [12], reasoning not about when to tell others, but rather when to ask for information. This re-casting is a conceptual contribution that we hope will stimulate a range of innovations, exploiting techniques produced in the time since work in [12] was first done, for instance, in learning-based and model-free approaches. It has led us to the method we detail, an algorithm most suitable for multi-agent systems operating under uncertainty, where broadcast communication is feasible but costly. In our method, agents integrate local observations and predict others' histories to form estimated joint beliefs, selecting actions to maximize expected rewards. Comparing outcomes with and without communication, they weigh the costs versus the benefits. Though relying on uncertain predictions can cause divergent actions, simulations show that tolerating some misalignment often outperforms overly conservative strategies. Beyond integrating with learning-based approaches, future work will focus on selective information sharing, where agents query only relevant teammates. This aims to improve scalability, support larger multi-robot systems, and enable testing in more challenging environments.

Acknowledgment

Part of the research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

References

[1] Y. Rizk, M. Awad, and E. W. Tunstel, "Decision making in multiagent systems: A survey," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 3, pp. 514–529, Sep. 2018.
[2] F. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*. Berlin, Germany: Springer, 2016.
[3] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artif. Intell.*, vol. 101, no. 1, pp. 99–134, 1998.
[4] C. Goldman and S. Zilberstein, "Optimizing information exchange in cooperative multi-agent systems," in *Proc. Int. Conf. Auton. Agents*, Jun. 2003, vol. 2, pp. 137–144.
[5] R. Nair, M. Roth, and M. Yokoo, "Communication for improving policy computation in distributed POMDPs," in *Proc. Int. Joint Conf. Auton. Agents Multiagent Syst.*, 2004, vol. 3, pp. 1098–1105.
[6] F. Wu, S. Zilberstein, and X. Chen, "Online planning for multi-agent systems with bounded communication," *Artif. Intell.*, vol. 175, no. 2, pp. 487–511, 2011.
[7] P. J. Gmytrasiewicz and P. Doshi, "A framework for sequential planning in multi-agent settings," *J. Artif. Intell. Res.*, vol. 24, pp. 49–79, Jul. 2005.
[8] P. Gmytrasiewicz, "How to do things with words: A Bayesian approach," *J. Artif. Intell. Res.*, vol. 68, pp. 753–776, Jul. 2020.
[9] L. Zhou and J. Luo, "A communication model for interactive POMDPs," in *Proc. Int. Conf. Comput. Sci. Educ.*, 2012, pp. 169–174.
[10] R. Emery-Montemerlo, "Game-theoretic control for robot teams," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, USA, 2005.
[11] C. Zhu, M. Dastani, and S. Wang, "A survey of multi-agent deep reinforcement learning with communication," in *Proc. 23rd Int. Conf. Auton. Agents Multiagent Syst.*, 2024, pp. 2845–2847.
[12] M. Roth, R. Simmons, and M. Veloso, "Reasoning about joint beliefs for execution-time communication decisions," in *Proc. Int. Joint Conf. Auton. Agents Multiagent Syst.*, 2005, pp. 786–793.
[13] S. Seuken and S. Zilberstein, "Formal models and algorithms for decentralized decision making under uncertainty," *Auton. Agents Multi-Agent Syst.*, vol. 17, no. 2, pp. 190–250, Oct. 2008.
[14] C. Goldman and S. Zilberstein, "Decentralized control of cooperative systems: Categorization and complexity analysis," *J. Artif. Intell. Res.*, vol. 22, pp. 143–174, Jun. 2011.
[15] H. Kurniawati, D. Hsu, and W. S. Lee, "SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces," in *Proc. Robot.: Sci. Syst.*, 2008, pp. 65–72.
[16] G. Shani, J. Pineau, and R. Kaplow, "A survey of point-based POMDP solvers," *Auton. Agents Multi-Agent Syst.*, vol. 27, pp. 1–51, 2012.
[17] M. Spaan and N. Vlassis, "Perseus: Randomized point-based value iteration for POMDPs," *J. Artif. Intell. Res.*, vol. 24, pp. 195–220, 2004.
[18] M. Roth, "Execution-time communication decisions for coordination of multi-agent teams," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, USA, 2007.
[19] M. Egorov, Z. N. Sunberg, E. Balaban, T. A. Wheeler, J. K. Gupta, and M. J. Kochenderfer, "POMDPs.jl: A framework for sequential decision making under uncertainty," *J. Mach. Learn. Res.*, vol. 18, no. 26, pp. 1–5, 2017.
[20] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella, "Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings," in *Proc. Int. Joint Conf. Artif. Intell.*, 2003, pp. 705–711.
[21] D. S. Bernstein, E. A. Hansen, and S. Zilberstein, "Bounded policy iteration for decentralized POMDPs," in *Proc. Int. Joint Conf. Artif. Intell.*, 2005, pp. 1287–1292.
[22] C. Amato and S. Zilberstein, "Achieving goals in decentralized POMDPs," in *Proc. Int. Joint Conf. Auton. Agents Multiagent Syst.*, 2009, pp. 593–600.
[23] J.-Y. Kwak, R. Yang, Z. Yin, M. E. Taylor, and M. Tambe, "Teamwork and coordination under model uncertainty in Dec-POMDPs," in *Proc. AAAI Conf. Interactive Decis. Theory Game Theory*, 2010, pp. 30–36.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement