Impact of Heterogeneity and Risk Aversion on Task Allocation in Multi-Agent Teams

Haochen Wu, Amin Ghadami, Alparslan Emrah Bayrak, Jonathon M. Smereka, and Bogdan I. Epureanu

***Abstract*—Cooperative multi-agent decision-making is a ubiquitous problem with many real-world applications. In many practical applications, it is desirable to design a multi-agent team with a heterogeneous composition where the agents can have different capabilities and levels of risk tolerance to address diverse requirements. While heterogeneity in multi-agent teams offers benefits, new challenges arise including how to find optimal heterogeneous team compositions and how to dynamically distribute tasks among agents in complex operations. In this work, we develop an artificial intelligence framework for multi-agent heterogeneous teams to dynamically learn task distributions among agents through reinforcement learning. The framework extends Decentralized Partially Observable Markov Decision Processes (Dec-POMDP) to be compatible to model various types of heterogeneity. We demonstrate our approach with a benchmark problem on a disaster relief scenario. The effect of heterogeneity and risk aversion in agent capabilities and decision-making strategies on the performance of multi-agent teams in uncertain environments is analyzed. Results show that a well-designed heterogeneous team outperforms its homogeneous counterpart and possesses higher adaptivity in uncertain environments.**

***Index Terms*—AI-Based methods, reinforcement learning, multi-robot systems, task planning, cooperating robots.**

I. Introduction

RECENT developments in autonomy offer opportunities that might lead to a paradigm shift in several domains. A multi-agent system design is beneficial in many aspects, particularly when a system is composed of multiple entities that are distributed functionally or spatially. Collaboration enables the agents to work as a team and complete activities that they are not able to accomplish individually. Instead of agents being centrally controlled, a decentralized multi-agent team can improve performance, robustness, and scalability by planning and performing actions in parallel.

Manuscript received February 24, 2021; accepted June 29, 2021. Date of publication July 14, 2021; date of current version July 28, 2021. This work was supported through the Automotive Research Center, University of Michigan, Ann Arbor, MI, USA. Distribution A: Approved for public release; distribution unlimited. OPSEC# 5021. This letter was recommended for publication by Associate Editor P. Ogren and Editor M. Vincze upon evaluation of the reviewers' comments. *(Corresponding author: Haochen Wu.)*

Haochen Wu, Amin Ghadami, and Bogdan I. Epureanu are with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: haochenw@umich.edu; aghadami@umich.edu; epureanu@umich.edu).

Alparslan Emrah Bayrak is with the School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail: ebayrak@stevens.edu).

Jonathon M. Smereka is with the US Army CCDC Ground Vehicle Systems Center, Warren, MI 48397 USA (e-mail: jonathon.m.smereka.civ@mail.mil).

Digital Object Identifier 10.1109/LRA.2021.3097259

To date, most work in multi-agent task allocation (MATA) [1] has focused on homogeneous team compositions [2] with the consideration of temporal constraints [3], communication protocols [4], and spatial dynamics of tasks [5]. Task allocation for heterogeneous teams [6]–[9] has also been addressed to study the collaborative behaviors of heterogeneous agents with specialized capabilities on handling various types of tasks. The solution approaches of MATA in literature include optimization-based methods [2], [3], [6]–[9] modelled as mixed integer programming problems and reinforcement learning (RL)-based methods [4], [5] which are formulated as Decentralized Partially Observable Markov Decision Processes (Dec-POMDP) [10] and solved by deep learning techniques [11]–[13] or heuristic search [14]. As computing systems continue to advance, there is a push towards considering more complex environments and diverse teams such as autonomous agents as team members alongside with humans in many real world applications [15], ranging from transportation systems to exploration of hazardous environments and rescue in disaster scenarios. Autonomous agents are capable of handling dangerous tasks but limited in reaction to unforeseen events. At the same time, humans have more adaptive and creative problem-solving skills but are limited in terms of handling specific tasks and cognitive loads. This inclusion of autonomy within a team context establishes a need to develop a framework to optimally train a team of heterogeneous agents and investigate how a team benefits from heterogeneity in an operation. The literature, however, lacks enhanced modeling of strategic task allocation for teams of heterogeneous agents that are managing dynamic task demands in urgent and uncertain operations. While heterogeneity in multi-agent teams may offer benefits, these benefits can only be understood by analyzing optimal heterogeneous team compositions and dynamically distributing tasks among agents in uncertain environments, which remain challenging. Although optimization-based MATA problems can handle constraints between heterogeneous tasks and agents [6], the task demands are usually satisfied upon one-step assignment [8] and analysis of dynamically changing demands affected by agent decisions and capabilities is still limited [7], [9]. In particular, a Dec-POMDP formulation has not been formally utilized by a team of heterogeneous agents to provide the comprehensive modeling of the dynamics between agent capabilities, task demands, and perception accuracy.

Unlike homogeneous teams where agents have the same functional capabilities and share the same parameters in their decision processes, heterogeneous team composition and training require a delicate effort considering the trade-off between

the stability and adaptive behavior of each agent. Heterogeneity in a team is often considered as the difference in rules of engagement or assignment constraints [6], task suitability [7], or functional heterogeneity [8]. In addition to the heterogeneity in agent task-related capabilities, the difference of agent risk tolerance in decision-making, which has rarely been mentioned in the literature of MATA, also plays a significant role in overall performance especially in human-autonomy teaming under uncertain environments. In certain occasions during an operation, an agent might take an action which is considered sub-optimal or risky contrasting the previous experience and training. Modeling risk aversion in the agent decision-making process allows assessing the value of the behavior in different situations. Therefore, this letter considers multiple sources of heterogeneity by varying agent capabilities and risk-averse factors in the agent decision-making process. The studies that explore the combined effect of multiple heterogeneity sources and provide parametric analysis of various heterogeneity factors on the team performance are essential for designing the next generation of multi-agent systems in real-world applications.

Here, we develop an artificial intelligence framework for multi-agent heterogeneous teams to dynamically learn task distributions among agents and maximize the performance in complex operations through reinforcement learning. The heterogeneity of a multi-agent team in this content is described by 1) the difference in agent capabilities of task handling, sensing, and communication that have direct impact on task level transitions and perceived information accuracy and 2) the difference in agent decision-making process representing the level of risk aversion. The framework extends Dec-POMDP [14] to the heterogeneous teaming where agents are equally responsible for strategic planning and execution, and the heterogeneity in agent capabilities and decision processes is explicitly modeled. The proposed approach employs deep learning techniques [11] to improve computational efficiency and utilizes belief representation to summarize past experience allowing the incorporation of risk tolerance in the framework. Based on the proposed approach, we analyze how different levels and types of heterogeneity in a team influence the team performance during an operation. In addition, we introduce and incorporate some characteristics of humans in a team as a basis to facilitate studying human-autonomy systems within this framework.

The contributions of this work are:

- Development of a decentralized task allocation framework for teams with heterogeneous agents, which enables complex teaming interactions in environments with dynamic demands and uncertainties
- Incorporation of risk aversion and perception accuracy in agent decision-making processes, and analysis of their effects on team performance in the presence of unforeseen events
- Quantitative investigation of the combined effects of heterogeneity and risk aversion on teaming performance

II. Methods

To formulate the multi-agent heterogeneous teaming problem, we assume that agents are equipped with their own decision-making processes and are equally responsible for strategic planning and execution. No centralized coordinator commanding team members is considered. In addition, to ensure heterogeneity, agents have the same type of attributes but different capability levels and decision models. The proposed method inherits decentralized reinforcement learning as demonstrated in Fig. 1. With functional capabilities to handle tasks, sensing capabilities to perceive information and communication capabilities to ensure information accuracy, all agents collaborate to achieve a common goal described by the rewarding system.

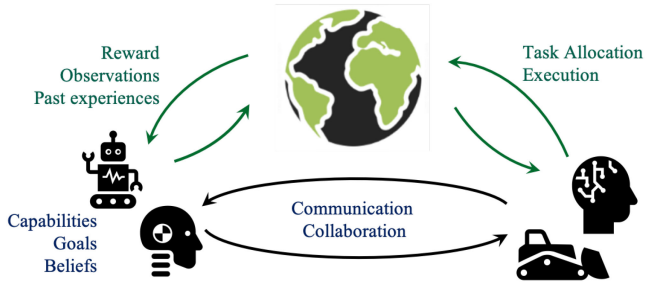


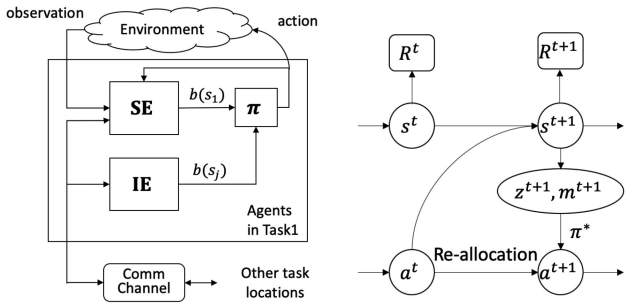
Fig. 1. Schematic of decentralized reinforcement learning approach in heterogeneous multi-agent task allocation problems.

A. Problem Formulation

The decision-making process for heterogeneous multi-agent teaming is formulated as an extended Dec-POMDP [14] where the capabilities of agents and information exchange are explicitly modeled. The proposed Heterogeneous Teaming Decentralized Partially Observable Markov Decision Process (HT Dec-POMDP) is defined as a tuple of $(G, \alpha, S, A, \mathcal{T}, R, Z, \mathcal{O}, M, \mathcal{I}, h, b^0, \gamma)$ where:

- $G := \{g_1, g_2, \ldots, g_p\}$ is a finite set of p tasks
- $\alpha := \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ is a finite set of n agents
- $S := S^D \times S^C, s \in S$, is the overall state that is factored into p tasks and described by task demand/severity levels $S^D := S_1^D \times S_2^D \times \cdots \times S_p^D$ and joint agent capabilities $S^C := S_1^C \times S_2^C \times \cdots \times S_p^C$
- $A := \times_i G, a \in A$, is the set of joint assignment decision by assigning α_i to g_j
- $\mathcal{T} := Pr(s'|s, a)$ is the state transition probability
- $R := f_R(s, a, s')$ is the reward function
- $Z := \times_i Z_i, z \in Z$, is the set of joint observations, where $z =< z_1, z_2, \ldots, z_n >$
- $\mathcal{O} := Pr(z|s,' a)$ is the observation probability
- $M := \times_i M_i, m \in M$, is the set of joint information, where M_i represents the state information collected by agent α_i
- $\mathcal{I} = Pr(m|s,' a, z)$ is the information probability
- h is the planning time horizon
- b^0 is the initial belief defined as the probabilistic distribution over the problem state space
- γ is the discount factor

HT Dec-POMDP provides a framework for heterogeneous teams to make decentralized decisions (Fig. 2(a)). The formulation splits the environment dynamics in the task domain, where the dynamics of each task can be individually modeled. However, the overall stochastic environment dynamics is maintained



(a) (b)

Fig. 2. (a) Belief update mechanism for agents. For an agent assigned to task 1, state estimator (SE) updates the belief $b(s_1)$, information estimator (IE) processes communicated observations and updates the belief over other tasks $b(s_j)$. The agent follows the policy π based on the overall belief updated for the next action. Comm Channel represents communication strategy which can range from all-to-all to no communication. (b) Demonstration of state transition by time steps in team view. Allocation decisions for the next time step a^{t+1} are made based on agents' observations z and communicated information m received after the previous action a^t

by considering conditional transition probabilities between tasks and task-specific capabilities, sensing-dependent observation probabilities, and communication-dependent information accuracy. Agents with higher sensing capabilities receive accurate information with higher probabilities, and the observations communicated by the agents with higher communication capabilities would be received accurately with higher probabilities. The state of the operation includes the severity level of each task S^D that evolves throughout the operation and the joint capability of the agents assigned to that task S^C depending on the pre-defined agent capabilities, i.e. $C_i = [c_1, \ldots, c_k]$ is the capability vector for agent α_i and k represents the number of attributes. The environment reacts to agents actions depending on the joint capability of agents assigned to each task. As shown in Fig. 2(b), during an operation, agents take actions a based on their belief over the system state and bring the state to another level in the next time step. Agents then receive partial observations z of task states through sensors, broadcast information m through limited communication, receive reward R for accomplishing the tasks. Agents further update their belief with collected information using state estimation (SE) and information estimation (IE) described in Fig. 2(a), and make decision for the next action under policy π using the collected information.

The primary objective of this formulation is to find optimal task allocation policy Π. Task allocation policy is defined as $\Pi : M \rightarrow A$, which maps the collected information M to decisions made by agents for the next action A. The optimal policy is defined as $\pi_* = argmax\{V(\pi)\}$ that maximizes team performance, where $V(\pi)$ represents the expected total future team reward defined as:

$$V_\pi = \mathbb{E}\left[\sum_{t=0}^{t=h} \gamma^t f_R(s^t, a^t) | \pi, b^0\right]. \tag{1}$$

Given the initial belief state b^0, the team follows the task allocation policy π for a time horizon of h steps. The reward is discounted by time (γ^t) reflecting that higher reward is given if the team can accomplish all tasks earlier in time. We propose a Q-learning approach to learn the policy based on the agents' beliefs as discussed in Section II-B.

B. Decentralized Deep Q-Learning With Beliefs

Dec-POMDP provides a decentralized decision model formulation but is known to be NEXPTIME-hard. The optimal policy could be found by mapping each agent action-observation history to an action, and the exact solution is only solvable in a finite time horizon [14]. In the context of multi-agent task allocation and the proposed HT Dec-POMDP framework, requiring agents to maintain information regarding all tasks thorough observations and communication makes the action-observation-information history very large, or even intractable, because a large time horizon is required usually for the team to complete the mission. The proposed approach extends the discrete state to continuous belief state space with a self-providing reward mechanism and allows agents to approximate the optimal solution in infinite time horizon. The mechanism of self-providing rewards allows agents to generate rewards based on their own belief states instead of relying on the reward provided by the environment, which facilitates modeling of heterogeneity in reward functions.

Belief state representation is helpful when solving a partially-observable environment [2], [14], [16]. A belief state is the probability distribution over the states of the environment identified by an agent. Instead of iterating all possible agent action-observation-information histories and to extend the solution to infinite time horizon, agent past experience is represented and summarized by a belief update mechanism. When the number of tasks and the number of task severity levels are large in complex operations, deep learning techniques could be used to learn the decisions based on beliefs directly.

$$\begin{aligned} b'(s') &= Pr(s'|m, z, a, b) \\ &= \frac{Pr(m|s,' a, z) Pr(s'|a, z, b)}{Pr(m|a, z, b)} \\ &= \frac{Pr(m|s,' a, z) Pr(z|s,' a) Pr(s'|a, b)}{Pr(m|a, z, b) Pr(z|a, b)} \\ &= \eta Pr(m|s,' a, z) Pr(z|s,' a) \sum_s Pr(s'|a, s) Pr(s|a) \\ &\propto \mathcal{I}(s,' a, z, m) \mathcal{O}(s,' a, z) \sum_s \mathcal{T}(s, a, s') b(s) \end{aligned} \tag{2}$$

Let $b(s)$ denote the probability of an agent α_i being in the state s. The belief update follows Bayes' rule and can be computed by the state transition, observation, and information probabilities $\mathcal{T}, \mathcal{O}, \mathcal{I}$ defined in Section II-A. The belief state inherits the Markov property since the updated belief does not depend on the previous belief states but only on the current state. Based on agent action a, observation z, and received information m, the updated belief state $b'(s')$ could be computed with Eq. 2, where $\eta = \frac{1}{Pr(m|a,z,b) Pr(z|a,b)}$ is considered as a normalizing factor which assures $\sum_{s'} b'(s') = 1$.

Being able to receive observations, communicate information and maintain extracted state information locally in the belief space, the agents could learn the optimal task distribution in a

Algorithm 1: Decentralized Deep Q-Learning With Beliefs.

1: Initialize replay memories D and Q-Networks with random weights *for all agents*
2: **for** each training episode **do**
3: Initialize operation and beliefs *for all agents*
4: **for** each operation step **do**
5: ***For all agents***
6: Make decisions with ϵ-greedy and execute decisions
7: Receive observations from the assigned tasks and communicate observations in **Comm Channel**
8: Update beliefs with **SE** and **IE**
9: Compute reward and store transition into D
10: Train Q-Network with a randomly sampled minibatch of transitions in D
11: **end for**
12: **end for**

completely decentralized fashion with the proposed HT Dec-POMDP formulation. The proposed deep Q-Learning approach shown in Algorithm 1 finds the optimal task allocation decision given the internal belief instead of being provided with the full observations of the environment [11]. The training algorithm is deployed to each agent, and agents learn the collaborative decision by interacting with the environment and storing past experience in the form of beliefs.

With processed observations and information in the state estimator and the information estimator (Fig. 2(a)), each agent updates the belief b as described in Eq. 3 (Alg. 1, Line 8), and the belief b summarizing the past observations and communicated information serves as the agent's state representation of the operation environment. Given a reward function directly on beliefs $f_R(b,a)$, the reward discount factor γ, and the time horizon h, the optimal action-value function $Q^*(b,a)$ is defined as the maximum possible expected total reward or Q-Value when having internal belief b and taking action $a=\pi(b)$ under the policy π as follows

$$Q^*(b,a) = \max_\pi \mathbb{E}\left[\sum_{t=t'}^{t=h} \gamma^{t-t'} f_R(b^t,a^t)|b^t=b, a^t=\pi(b)\right]. \tag{3}$$

Then, the optimal policy can be retrieved by $\pi^*(b) = \text{argmax}_a Q^*(b,a)$. The optimal Q-Value $Q^*(b,a)$ has to satisfy the *Bellman Optimality* to maximize the expected Q-Value of the next belief-action pair $(b,' a')$:

$$Q^*(b,a) = \mathbb{E}_{b'}[f_R(b,a) + \gamma \text{max} Q^*(b,' a')|b,a]. \tag{4}$$

As shown in Fig. 3, we use a neural network to approximate the Q-Value $Q(b,a;\omega)$ with belief states as network inputs and train the Q-Network for each agent so that $Q(b,a;\omega) \approx Q^*(b,a)$. The Q-Network takes belief state as input and outputs the Q-Value corresponding to each action. Given transitions $< b,a,b,' f_R(b,a) >$ (Alg. 1, Line 10), the Q-Network is trained by minimizing the loss function $L_i(\omega_i)$ between predicted Q-Value $Q(b,a;\omega_i)$ and the target Q-Value g_i for each iteration i, where $g_i = f_R(b,a) + \gamma \text{max} Q(b,' a';\hat{\omega}_i)$ and $\hat{\omega}_i$ is the target network parameter which is updated less frequently:

$$L_i(\omega_i) = \mathbb{E}_{b,a,f_R,b'}[g_i(b,' a,' \hat{\omega}_i) - Q(b,a;\omega_i)]. \tag{5}$$

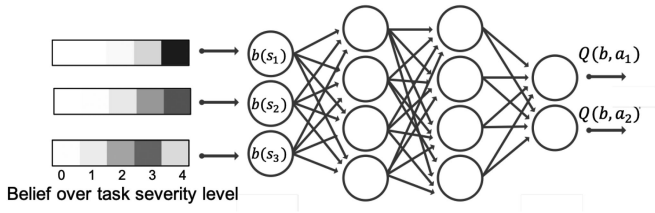


Fig. 3. Schematic of a Q-Network with an agent's belief over severity levels of 3 tasks as input and Q-Values of 2 actions as output. In the proposed approach, agent's belief over task severity levels are used as an input to the Q-Network and the outputs are the Q-Values in the action space.

The choice of reward function is an important factor in the reinforcement learning process, and finding optimal reward function can expedite the training process [17], [18]. In our task allocation framework, agents receive an ultimate high reward upon mission completion (all task severity levels are zero), while a comparably small reward is given to agents for partially completing the mission (only a fraction of tasks are completed). This reward structure can incentivize agents to reach mission completion during training. In particular, we define the reward function based on the agent's belief as:

$$f_R(b) = W_l^T R_l \tag{6}$$

The level completion reward R_l is the mission goal reward describing how much reward is received at each task severity level l. The level completion weight is defined as a function of belief with two tuning parameters: $W_l = f(b;w,b_{th})$, where $w \in [0,1]$ is the weight for partial completion, representing how much agents reward themselves with partial credits and $b_{th} \in [0,1]$ is belief reward threshold, deciding how generous agents reward themselves since some agents want to be more sure about the situation in order to reward themselves. Here, agents receive up to 30% of the maximum reward when up to 75% of the tasks are completed. They receive 100% of the reward upon the completion of all the tasks. In addition, the effect of variation in the belief reward threshold as a risk factor in the decision process is investigated in detail in Section III-B2.

III. Results

We showcase our analysis on a disaster relief scenario which is an extension of the Firefighting domain, a benchmark used in the evaluation of Dec-POMDP planning algorithms [14]. In this new scenario, a team of n agents have to extinguish fires in p houses and safely rescue people trapped in the burning houses. The fire extinguishing and rescue demands can vary in the range of [0,4], where 4 is the worst case scenario and zero represents that the task is completed. Each agent has its unique capabilities in extinguishing fire and rescuing people. In particular, we define four attributes for each agent, namely fire extinguishing, rescue, sensing, and communication capabilities. We consider the capability level for each attribute as a discrete value in the range of [0, 5], where higher values represent better agent capability and zero represents that the agent is incapable

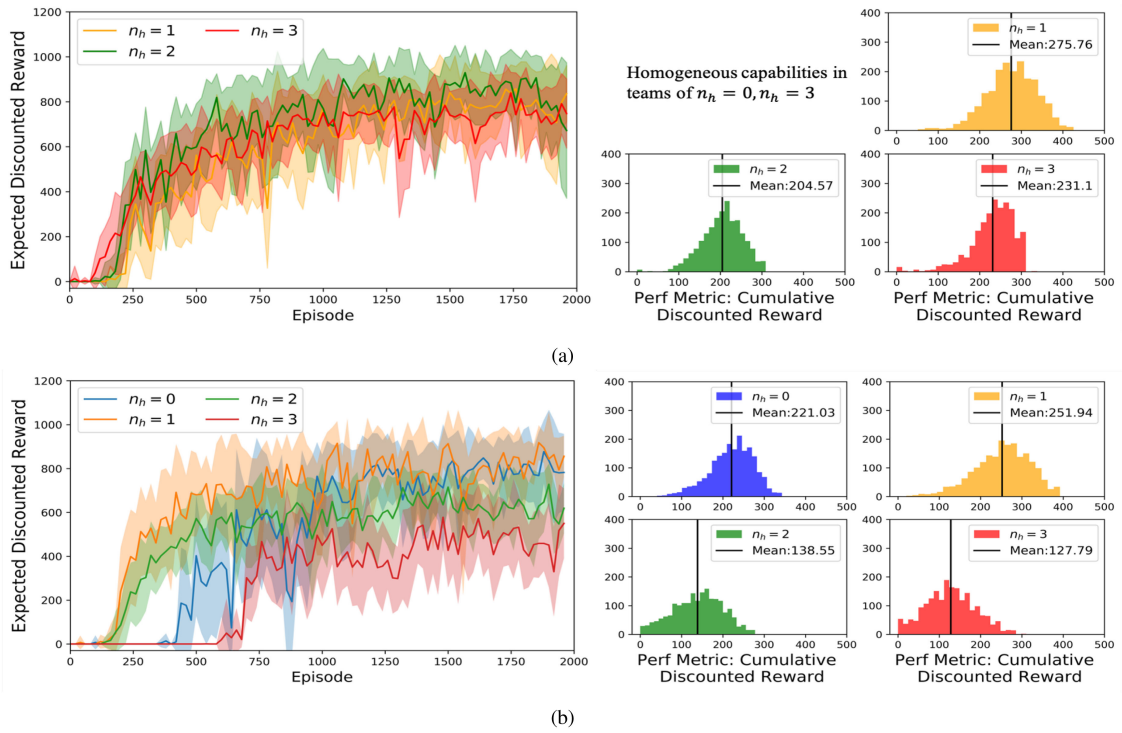


Fig. 4. Learning curves with shaded area indicating standard deviations of every 20 training episodes and performance evaluations with 2000 runs for different team compositions with (a) heterogeneity in capabilities only (b) heterogeneity in capabilities and decision models

in an attribute. The higher the joint capability of the agents assigned to the same task is, the higher is the probability of the demand at that task to be reduced, which implies agents have to collaborate to resolve the worst situation with the highest probability. Higher sensing and communication capabilities help the agents to approximate the environment status with higher accuracy. Agents would learn to cooperate on the same task when task demand levels are high and synergistically work on different tasks at the same time when task demand levels are low to resolve the severe situation faster. They might also decide to stay at certain locations to prevent the situation from getting worse.

A. *Effect of Heterogeneity on Team Performance*

1) Heterogeneity in Capabilities: In a heterogeneous team, agents may be specialized and have different capability levels. To study the effects of heterogeneity in agent capabilities on the team performance, we consider different team compositions. To ensure the fairness on team capabilities, the mean capability levels for firefighting and rescue attributes are kept at 2. For the same attribute, when there is an agent with lower capability, there is another agent in the same team with higher capability to keep the mean constant. In order to focus on heterogeneity of task-specific capabilities, the sensing capability and communication capability of all agents in all teams are fixed at level 2. With a team of three agents and four attributes (i.e. firefighting, rescue, sensing, communication) mentioned above, there are four possible team compositions represented by $n_h = 0, 1, 2, 3$, where n_h represents the number of agents with lower than average capabilities. For instance, in the case of $n_h = 1$, one agent has firefighting capability 1 and rescue capability 1, and the other two agents have capabilities ≥ 2 to keep the mean capability of each attribute as 2. Note that both $n_h = 0$ and $n_h = 3$ teams have homogeneous capabilities where $C_i = [2, 2, 2, 2]$ for all agents in order to satisfy the above conditions.

Fig. 4(a) shows the learning curves and performance evaluations for team compositions with different capability distributions assuming all agents decide rationally. After training, the performance of each team is evaluated in 2000 operation trials shown in the histograms. Performance metric is the discounted reward/score given by the environment. This metric indicates how effectively a team completes all tasks in execution. Higher performance indicates the team is able to reduce all task severity levels to zero within fewer time steps. Note that the reward during training is obtained using agent belief, which is different than the performance metric, but it is still reflecting the team performance. Results show that heterogeneity in agents' capabilities affects the overall performance of the team. In the simulated experiments, for the heterogeneous team of $n_h = 1$ the team has the best performance based on the mean cumulative discounted reward achieved by the team in 2000 evaluations. When agents have specialized capabilities ($n_h = 2$) or uniform capabilities ($n_h = 0$ or 3), the team performance deteriorates.

2) Heterogeneity in Capabilities and Decision Models: Next, we study the effect of sub-optimal decision-makers on the performance of studied team compositions of the previous example. Agents and in particular humans in the team might not always make the optimal (rational) decision since sometimes they are risk takers and tend to make sub-optimal decisions. This behavior is called noisy rational decision, and it could represent human creativity in learning and execution. In the

previous example, we assume that the agents with relatively lower capabilities (represented by n_h) would make less rational decisions resulting in heterogeneity in decision-making across the team. We consider these agents as human-like agents assuming that humans have less task handling capabilities but their creativity and higher risk tolerance leads to noisy rational decisions. To model noise in agent decisions, we incorporate a noisy rational model, a widely used human decision model in cognitive science [19]–[21], into our proposed framework. In particular, we give an agent the option to take any action with certain probability defined as

$$Pr(a|b) = \frac{exp(\theta \cdot Q(b,a))}{\sum_{a_i \in A_i} exp(\theta \cdot Q(b,a_i))}, \tag{7}$$

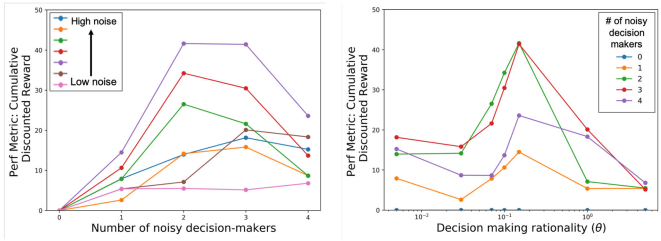
where A_i denotes the action space for an agent. The noisy rational model takes the output of Q-Network (Q-Value $Q(b,a)$ described in Section II-B) and computes the the probability of taking each possible action a given a belief state b. The tuning parameter $\theta \in [0, \inf]$ determines how rational the agent makes the decision. When $\theta = 0$, agents make completely random decisions; when $\theta \to \inf$, and agent makes rational decision, i.e., chooses the action with the highest Q-Value based on its Q-Network. A noisy decision is sub-optimal with the trained network but it could be a better decision if training is insufficient.

With added noisy rational decisions to agents with lower than average capabilities, we observe that, in the training curves in Fig. 4(b), the teams with homogeneous capabilities (i.e., $n_h = 0$ and $n_h = 3$) start to get reward much later than the other teams, implying that homogeneous teams or a team where all its agents often make sub-optimal decisions require longer training. In the performance evaluation, the team with $n_h = 1$ still has the best performance among all team compositions highlighting that forming specialized or a uniform team does not lead to the best training efficiency and performance.

Results of the analysis in this Section show that how a heterogeneous team can outperform a homogeneous one in complex operation scenarios. The proposed framework is capable of capturing the deficiencies in the training for different teams and the uncertainties in the team performance and provides insights about optimal selection of a team of agents leading to stable training and reliable performance.

B. *Effect of Risk Aversion on Team Performance*

1) Auxiliary Tasks: It is often challenging to deploy the trained agents in a real operation where they face events that they are not trained for. We define auxiliary tasks to simulate unexpected events during operation, i.e. tasks that are present in operation scenario but were not present in training. In circumstances where agents encounter auxiliary tasks, the proportion of sub-optimal decision-makers in the team as well as the level of noise in the agent's decisions can affect the team performance significantly. Here, we evaluate the effect of facing risks in agents' decision-making process when tackling this unexpected situation. Such a scenario is analogous to a team including humans who have different level of creativity and/or uncertainty in their decisions.



(a) (b)

Fig. 5. (a) Effect of number of noisy decision-makers in a team on the team performance, (b) effect of decision-making rationality on the team performance. Each point represents the mean value of 1000 evaluations.

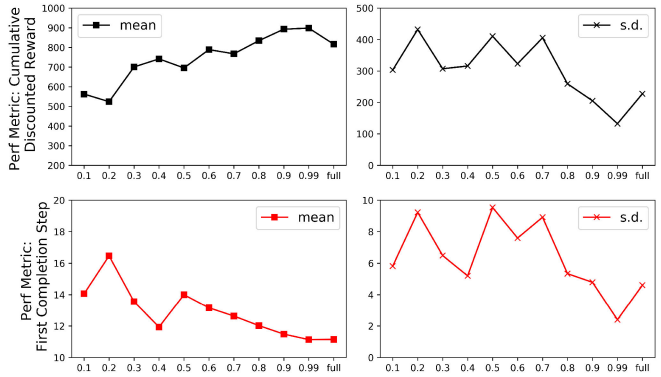


Fig. 6. Effect of belief reward threshold (b_{th}) on team performance, i.e., cumulative reward (top row) and first completion step of all tasks (bottom row). Results are compared to cases when system provides reward on full completion (full) of the tasks. The mean value and their corresponding standard deviations are obtained from 1000 simulations for each case.

In this study, we explored the effect of heterogeneity in decision-making over the team members on completing a mission including tasks which were not present during the training (i.e., auxiliary tasks). We considered 6 regular tasks, 4 agents, and added 2 auxiliary tasks in training for exploring the ability of the team to adapt to unforeseen demands in performance evaluation. Auxiliary tasks remain at zero level during training and start at the highest level in evaluation. Fig. 5(a) shows the performance of different team compositions when the number of noisy (sub-optimal) decision-makers in the team is varied. In addition, Fig. 5(b) shows the performance of each team by varying the level of noise in the agents' decision-making capabilities. Results show that there is a trade-off between the number of noisy decision-makers and level of rationality. In particular, it is observed that there exists an optimal value for both the number of noisy decision-makers in the team and the level of noise in their decision-making processes which result in maximum reward and the best performance. A carefully designed team including risk taker agents enables the team to handle unforeseen circumstances and make up for the deficiencies of the Q-Network by occasionally working on auxiliary tasks or tasks assumed to be sub-optimal choices under the Q-Network.

2) Risk Tolerance/Award Generosity: In a decentralized decision-making process, agents obtain rewards based on their evaluation of the environment status after each action. Such a judgment is made based on agents' belief identifying how

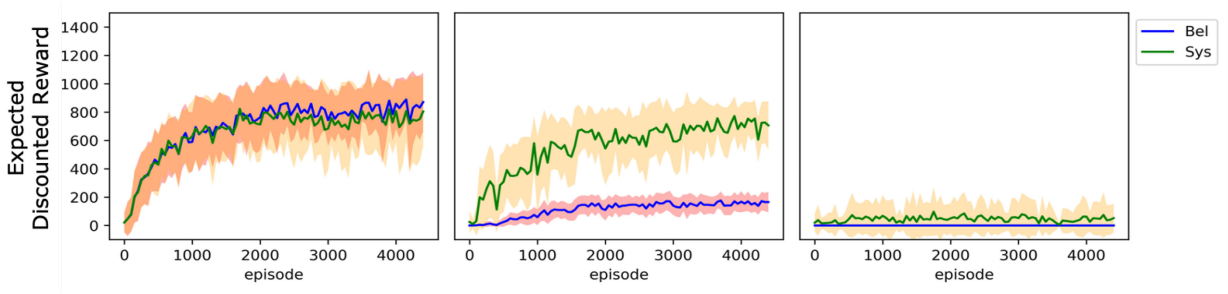


Fig. 7. Low (left), medium (center), and high (right) levels of added uniform prior in training. Blue lines indicate average agent belief reward and green lines indicate the corresponding reward system would have given. Solid lines and shaded area represent the mean value and standard deviations of every 20 episodes during training.

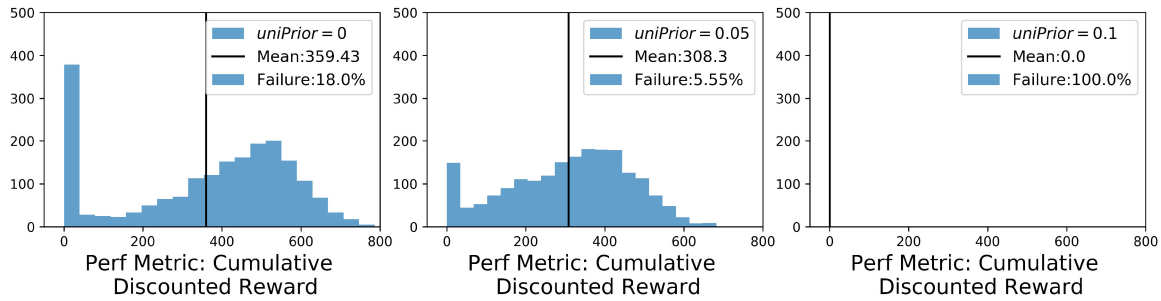


Fig. 8. Low (left), medium (center), and high (right) levels of added uniform prior in performance evaluation with a sudden event.

certain an agent thinks about that a task is completed. In a heterogeneous team, different agents (e.g. humans and autonomous agents) might have different levels of standard to label a task as completed. Here, we investigate the effect of the belief threshold, which is another risk factor in decision-making process, on the team performance.

We considered a team of three agents with similar capabilities in the disaster relief scenario. The belief reward threshold of the agents are varied from 10% to 99% and the team performance is evaluated for each case. A $p\%$ belief threshold means that the agent considers a task to be complete if it is at least $p\%$ certain about that. An agent considers a mission completed when its belief over completion of all the demands (i.e., fire and rescue levels at all locations) falls within its belief threshold. The performance evaluation includes 1000 realizations with random but heavy (≥ 3) initial demand levels. We investigate two performance metrics; the cumulative discounted reward and the first completion step of all tasks (Fig. 6).

Results show that the belief reward threshold of $b_{th} = 0.9, 0.99$ lead to the best performance based on received cumulative discounted reward. The performance outperforms the case of system reward, i.e., when the reward is given to the agents by system during learning instead of their belief over environment status. In addition, there is trade-off in the belief threshold. With a large threshold, agents receive reward for accurate state inference and therefore significantly improve the performance uncertainty represented by standard deviations of the evaluated rewards. With small thresholds, however, agents receive reward more frequently to help learning, but the performance uncertainty is significantly larger than high belief thresholds, meaning that it is risky to select an agent which is generously rewarding itself in the team and the performance uncertainty is not guaranteed.

3) Balancing Past Experience and Recent Observations: In real operations, a team of agents usually encounters sudden events that have not been experienced during training. Making decisions purely based on past experience leads to reluctant behaviors in reaction to sudden events. The incapability of handling these situations might result in failure in completing the defined tasks for the agents. Inspired from human adaptivity and the ability to promptly assess the current environment status, the agent's ability of balancing past experience and recent observations is modeled in the developed framework. In particular, we define a parameter u called uniform prior that adjusts belief probability distributions before receiving observation as a tuning parameter. Such a parameter identifies to what extent an agent takes the risk of trusting recent observations rather than the past experience in its decisions for the next action.

Figure 7 shows the training curves with three levels of added uniform prior. When a uniform prior of value $u \in [0, 1]$ is high and is added to the belief probability, by re-normalizing the probability, the agent belief gets proportionally biased toward u before taking observation and communication information. Higher u indicates and agent would rely more on the upcoming information to make decision. Results show that there is a trade-off between the effect of current observations and past experience on decision learning and performance of a team. A higher prior, when agents base their decisions more on recent observations, results in the worst training quality in regular operations. In Fig. 8, trained agents are evaluated in an operation with a sudden event which is not present during the training. Particularly, we increase the fire levels of the houses to their

maximum level after ten time steps since the start of the simulations. The agents are expected to complete the tasks, i.e., bring all fire levels to zero and rescue all people, in 30 time steps; otherwise, the operation is considered as failed. We observed that a medium level of uniform prior could help the team react to such sudden events. The failure rate of completing all tasks is noticeably lower than the case when agents base their decisions purely on past experience or recent observations.

IV. Conclusion

In this letter, we developed an artificial intelligence framework for multi-agent heterogeneous teams to dynamically learn task distribution and maximize the performance in complex operations through reinforcement learning. The proposed framework HT Dec-POMDP is compatible to model various sources of heterogeneity within a team, captures aspects of intelligence that produce collaborative teaming, and provides the opportunity to quantitatively investigate the effects of heterogeneity and risk aversion on task allocation in multi-agent systems. Results of this study show that a well-designed heterogeneous team significantly improves the team performance, and including agents with higher risk tolerance in a team significantly improves the chance of achievement in uncertain environments.

The proposed framework can be used as a basis for further developments and design of multi-agent systems, particularly human-autonomy teams as one of the future works, where forming a team composed of agents with heterogeneous capabilities and different levels of creativity and risk tolerance is inevitable. Note that improving the computational efficiency of existing algorithms might be required particularly when more types of tasks and larger teams are involved. Nevertheless, the proposed approach facilitates accounting for information disruption and risky behaviors, and can provide insights into making high-level decisions in complicated operations. Moreover, with the help of deep learning approaches and defined reward functions, additional types of information can be included into the belief, as the operational scenario may require. Examples include information regarding previous assignments and consideration of cost of traversing between task locations for spatially distributed demands.

References

[1] S. Saravanan, K. C. Ramanathan, R. MM, and M. N. Janardhanan, "Review on state-of-the-art dynamic task allocation strategies for multiple-robot systems," *Ind. Robot*, vol. 47, pp. 929–942, 2020.

[2] R. Nair, M. Tambe, M. Yokoo, D. V. Pynadath, and S. Marsella, "Taming decentralized Pomdps: Towards efficient policy computation for multiagent settings," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 705–711.

[3] S. Amador, S. Okamoto, and R. Zivan, "Dynamic multi-agent task allocation with spatial and temporal constraints," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2014, pp. 1495–1496.

[4] D. B. Noureddine, A. Gharbi, and S. Ahmed, "Multi-agent deep reinforcement learning for task allocation in dynamic environment," in *Proc. Int. Conf. Softw. Technol.*, 2017, pp. 17–26.

[5] S. Omidshafiei, J. Pazis, C. Amato, J. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 2681–2690.

[6] H. Choi, A. K. Whitten, and J. P. How, "Decentralized task allocation for heterogeneous teams with cooperation constraints," in *Proc. Amer. Control Conf.*, 2010, pp. 3057–3062.

[7] Y. Emam, S. Mayya, G. Notomista, A. Bohannon, and M. Egerstedt, "Adaptive task allocation for heterogeneous multi-robot teams with evolving and unknown robot capabilities," in *Proc. Int. Conf. Robot. Automat.*, 2020, pp. 7719–7725.

[8] B. Fu, W. Smith, D. Rizzo, M. Castanier, and K. Barton, "Heterogeneous vehicle routing and teaming with gaussian distributed energy uncertainty," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 4315–4322.

[9] G. Notomista, S. Mayya, S. Hutchinson, and M. Egerstedt, "An optimal task allocation strategy for heterogeneous multi-robot systems," in *Proc. Eur. Control Conf.*, 2019, pp. 2071–2076.

[10] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," in *Proc. 16th Conf. Uncertainty Artif. Intell.*, vol. 27, no. 4, 2002, pp. 32–37.

[11] V. Mnih *et al.*, "Playing Atari with deep reinforcement learning," in *Proc. Neural Inf. Process. Syst.*, CoRR vol. abs/1312.5602, 2013, pp. 1–9.

[12] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable MDPS," in *Proc. Assoc. Adv. Artif. Intell.*, CoRR vol. abs/1507.06527, 2015, pp. 1–7.

[13] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Mach. Learn.*, vol. 8, no. 3/4, pp. 293–321, 1992.

[14] F. A. Oliehoek, M. T. J. Spaan, and N. A. Vlassis, "Optimal and approximate q-value functions for decentralized POMDPs," *J. Artif. Intell. Res.*, vol. 32, pp. 289–353, 2011.

[15] J. Y. C. Chen and M. J. Barnes, "Human-agent teaming for Multirobot control: A review of human factors issues," *Trans. Human-Mach. Syst.*, vol. 44, no. 1, pp. 13–29, 2014.

[16] 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.

[17] B. Liu, S. Singh, R. Lewis, and S. Qin, "Optimal rewards in multiagent teams," in *Proc. Int. Conf. Develop. Learn. Epigenetic Robot.*, 2012, pp. 1–8.

[18] Y. Dong, X. Tang, and Y. Yuan, "Principled reward shaping for reinforcement learning via lyapunov stability theory," *Neurocomputing*, vol. 393, pp. 83–90, 2020.

[19] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis. Courier Corporation*, North Chelmsford, MA, USA, 2012.

[20] R. Holladay, S. Javdani, A. Dragan, and S. Srinivasa, "Active comparison based learning incorporating user uncertainty and noise," *RSS Workshop on Model Learn. Human-Robot Commun.*, 2016.

[21] C. Basu, E. Biyik, Z. He, M. Singhal, and D. Sadigh, "Active learning of reward dynamics from hierarchical queries," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 120–127.