Effective Task Allocation in Ad Hoc Human-Agent Teams (Full)

Sami Abuhaimed and Sandip Sen

Tandy School of Computer Science, The University of Tulsa
{saa8061,sandip}@utulsa.edu

Abstract. With accelerated progress in autonomous agent capabilities, mixed human and agent teams will become increasingly commonplace in both our personal and professional spheres. Hence, further examination of factors affecting coordination efficacy in these types of teams are needed to inform the design and use of effective human-agent teams. Ad hoc human-agent teams, where team members interact without prior experience with teammates and only for a limited number of interactions, will be commonplace in dynamic environments with short opportunity windows for coordination between diverse groups. We study virtual ad-hoc team scenarios pairing a human with an agent where both need to assess and adapt to the capabilities of the partner to maximize team performance. In this work, we investigate the relative efficacy of two human-agent coordination protocols that differ in the team member responsible for allocating tasks to the team. We designed, implemented, and experimented with an environment in which virtual human-agent teams repeatedly coordinate to complete heterogeneous task sets.

Keywords: Human-agent coordination · Team performance · Task allocation

1 Introduction

Recent intelligent agent applications assume traditionally human roles in human-agent teams, e.g., tutor [34] and trainer [22]. Agents can also coordinate with people in critical tasks, including guiding emergency evacuations [32] and disaster relief [31]. New environments have been developed recently to enable group activities or coordination between people and agents, such as crowd-work and multiplayer online games. Human and agent teams are increasingly commonplace where they play different team roles. Since human-agent teams are being recognized as a routine and functionally critical important component of our societies, researchers have been studying the interactions and dynamics within these teams to understand and improve on their design [13]. Such human-agent teams have been studied in physical (robotic) and virtual settings [33].

We are studying ad hoc coordination scenarios where humans start coordinating with agents in a new environment with no prior interaction experience with the agent. The agent also does not have prior knowledge about its human

partners' abilities and preferences. Such coordination environments correspond to *ad hoc teams*: *An ad hoc team setting is one in which teammates must work together to obtain a common goal, but without any prior agreement regarding how to work together* [11]. Coordination in ad hoc teams is more challenging because of absence of prior knowledge and established relationships. Ad hoc human-agent coordination also raises critical new issues compared to ad hoc agent teams.

In this paper, we consider ad hoc teams trying to accomplish a set of tasks chosen from diverse task types. We assume that different human users will have different competence and expertise over various task types. We use a fixed agent expertise distribution (simulated) over the task types. To optimize the performance of a given human-agent team, therefore, it is necessary to have different task allocation distributions to the team members based on the expertise of the human team member. The allocation problem is exacerbated by the fact that a team member does not know the expertise levels of its partner *a priori*. While we allow for human and agent partners to share their estimated expertise over different task types, the accuracy and consistency of such expressed estimates by humans are unreliable [17].

Repeated interaction allows partners to refine the initial estimates provided, but such opportunities are few due to (i) only a limited number of repeated teamwork episodes and (ii) allocation decisions that determine what task types are performed by a partner in an episode. The success of such ad hoc human-agent teams in completing assigned team tasks, therefore, will critically depend on effective adaptability in the task allocation process.

Task allocation have been studied extensively in agent teams [26] as well as in human team and organizations literature [30]. However, we are not aware of prior examination of autonomous agents with task allocation roles, compared to humans, in virtual and ad hoc human-agent teams.

Some critical questions on task allocation decisions and human-agent ad hoc team efficacy that we study in this paper are:
• Is the performance of human-agent teams influenced by who allocates the tasks? If so, who produce higher team performance?
• How is the performance of human-agent teams affected by over/under-confidence of humans in their performance on different task types?
• How quickly can the task allocator in an ad hoc human-agent team learn about the relative capabilities of team members to optimize allocation of tasks?

We designed a new human-agent team coordination framework for task allocation and performance analysis: the Collaborative Human-Agent Taskboard (CHATboard). We use CHATboard for ad hoc human-agent team coordination, for repeated team task allocation scenarios, with human workers recruited from the Amazon Mechanical Turk (MTurk) platform. We present some conjectures as hypotheses about human confidence level in their expertise, about the relative effectiveness of human and agent task allocators, about the ability of agents to learn about human capabilities and adapt task allocations, and the ability of agents to harness human potential. We ran experiments involving repeated coordination using the Human and Agent Allocation protocols. We present the

results and our analysis to confirm our hypotheses and identify interesting phenomena that suggests future research tasks.

2 Related Work

Human-agent teams have been studied in different domains such as space robotics [13], therapy [1], deception-detection [20], programming [23] and decision-making [3]. The focus has been on agents who play supportive roles to human teammates [20], and they have been studied in robotic and simulation settings [33].

We, however, focus on an ad hoc environment, whereas studies, such as [13], incorporate training or interaction sessions with the agent and environment prior to the study. We are also interested in agents that are autonomous; DeChurch and Larson view an autonomous agent as a "team member fulfilling a distinct role in the team and making a unique contribution" [21].

Task allocation has been studied extensively in multi-agent teams [18, 12, 26, 14, 27]. In agent teams, the focus is on designing efficient mechanisms for agents to distribute tasks within their society; current approaches include integer programming [9], genetic [28], consensus and auction algorithms [6], and markets [8], and in domains such as Search and Rescue [36]. There is a recent focus on ad hoc environments [5] in which agents coordinate without pre-coordination. The majority of agent teams work is focused on simulation and robotic environments, and few have studied task allocation in ad hoc human-agent teams. Moreover, there is a general lack of investigating environments that include human teammates; including humans in same agent teams may require new approaches, as we do not know If the same mechanisms would produce similar results.

Task allocation is also studied in humans' team and organization literature. The mechanism of task allocation, which includes capabilities identification, role specification, and task planning, is considered an important component of teamwork [25, 24, 10]. Any organization needs to solve four universal problems, including task allocation, to achieve its goals [30]. In human teams, the focus is on understanding human team characteristics to design the best possible task allocation mechanism; however, there is little investigation of autonomous agents' effects on human teams when they are included in teams' allocation mechanisms.

Thus, the study of task allocation with combined human and agent team members is promising [4, 33]. The few existing work examine different dimensions. [33] and [31] investigate an agent assisting humans' control of robots in a simulation and experiments; the focus is supporting operators. Some of this work do not empirically investigate the area, focused on industrial settings, configure the agent in supporting roles, and it is unclear whether human participants received training prior to experiments, which means that the scenario not ad hoc.

In summary, studies that investigate task allocation within teams composed of humans and autonomous agents in ad hoc environments over repeated interactions are limited. We, therefore, study task allocation in ad hoc human-agent teams while being informed by potential human miscalibration tendencies.

3 Hypotheses development

We now motivate and present a number of research hypotheses related to ad hoc human-agent team task allocation and team performance that we will be experimentally evaluating in this paper. We study two task allocation protocols that govern the human-agent teamwork: Human Allocator Protocol and Agent Allocator Protocol. The former assigns task allocator role to human teammate, and the later to agent teammate (Section 5 presents more details).

We assume that there is considerable variability in the ability to complete routine tasks amongst average citizens. If this was not the case, human expertise in tasks can be gauged offline, and optimal task allocation can be performed, i.e., ad hoc teams would be no different than teams with significant prior working experience.

Hypothesis 0a (H0a)*: Different human participant has different perception and actual performance on different task types.*

We also assume that humans are unable to accurately estimate or express their performance (confidence levels) on different, somewhat routine task types. If this was not the case, then again, we could simply ask the human about their expertise levels for different task types and use that accurate information for task allocation, i.e., ad hoc teams would be no different than teams with significant prior working experience.

Hypothesis 0b (H0b)*: Human's average confidence levels on task types are not consistent with their performance on those task types.* We conjecture that the agent allocator has several advantages over the human allocator for effectively allocating team tasks: (a) lack of personal bias or preference for task types that is not performance motivated (for example, humans may like to do certain tasks even though they may not be good at it), (b) agents will have better estimates of their capabilities on known task types whereas humans typically over or under-estimate their expertise or performance on task types, (c) agents can consistently follow optimal allocation procedures given confidence levels over task types, (d) agents can more consistently learn from task performance of teammates in early episodes to update confidence level estimates and adapt task allocation to improve performance. This lack of bias may also result in the agent allocator allocating tasks such that together with higher team performance we also observe better performance of the human team member, i.e., better realize the human potential, compared to when the humans allocate tasks between team members! These conjectures are reflected in the following set of hypotheses:

Hypothesis 1 (H1)*: Agent Allocator Protocol produces higher teamwork overall performance than Human Allocator Protocol.*

Hypothesis 2 (H2)*: Agent Allocator can learn from ad hoc teamwork experience to quickly improve team performance through adaptation.*

Hypothesis 3 (H3)*: Agent allocator will engender higher Human potential realization compared to the Human Allocator.*

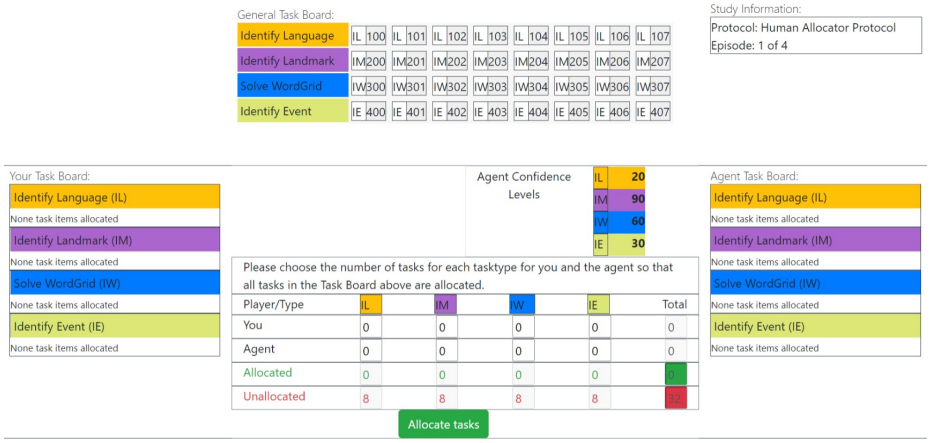

Fig. 1. CHATboard showing allocation phase of Human Allocation Protocol.

4 Collaborative Human-Agent Taskboard (CHATboard)

For systematic experimentation to evaluate the above hypotheses, we needed a domain that encapsulates the following characteristics:

• The team tasks used should be such that there would be significant variation in expertise level in the general populace. Larger variability would allow for more space for team adaptation and for human satisfaction with teamwork. We should also have the latitude to easily and believably configure varying agent capability distribution over the task types.

• The domain should allow an agent to be perceived as autonomous and playing a distinct peer role in the team.

• The domain should not require significant prior knowledge or training for human participants and should be accessible to non-experts for effectively operating in an ad hoc team setting.

• There should be flexibility in sharing team information, including task allocations and completions, with team members. The environment should be configurable between perfect and imperfect information scenarios as necessitated by the research question being investigated.

We developed CHATboard, an environment that facilitates human-agent, as well as human-human, team coordination. CHATboard contains a graphical interface that supports human-agent team coordination to complete a set of tasks (see Figure 1). CHATboard allows for displaying the task sets to be completed, supports multiple task allocation protocols, communication between team members for expressing confidence levels, displaying task allocations and performance by team members on assigned tasks, etc.

The framework utilizes the concept of tasks posted on blackboards, often used in coordination within human teams, to facilitate a human team member perceiving an agent as a distinct team member. Blackboards have also been effectively used in agent teams as a common repository for information sharing

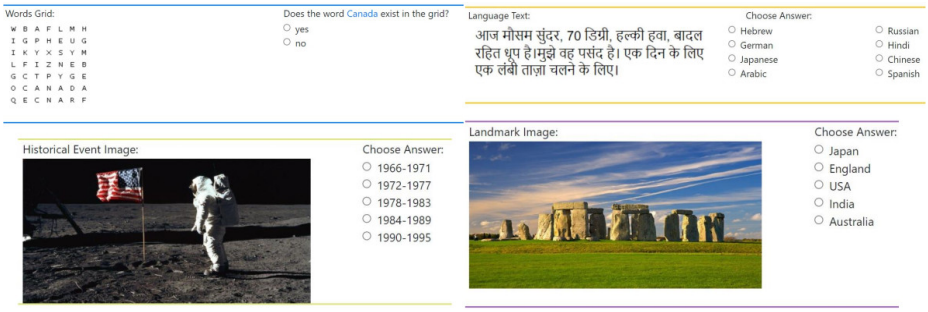

Fig. 2. Instances of different task types.

between agents [16]. Figure 1 shows the shared taskboad on top, which includes the set of team tasks organized by type, and two other boards respectively for the tasks assigned to the human and the agent team member. Figure 2 presents examples of task types. These task boards facilitate coordination, and act as easily navigable repositories for team information allowing team members to share and view information through these boards.

We define a set of n team members N: $\{p_1, p_2,...,p_n\}$, a set of m task types M: $\{y_1, y_2,...,y_m\}$, a set of r tasks, T_{jr}:$\{t_{j1}, t_{j2},...,t_{jr}\}$, for each task type y_j. Team member i can share their confidence levels $p_i(y_j)$ over task types y_j. The set C_i:$\{p_i(y_1),p_i(y_2),....,p_i(y_m)\}$ represent confidence levels for different task types for team player, p_i. The team members will interact over E episodes, where episode numbers range from $1 \ldots E$. $A_{i,e}$ denotes the set of tasks allocated to player i in episode e and we assume that all available tasks are exhaustively allocated, i.e., $\bigcup_i A_{i,e} = \bigcup_j T_{jr}$. The performance of player p_i for a task t_{jk} in episode e is referred to as $o_{ijke} \in \{0, 1\}$. We define the performance of p_i on task type y_j in episode e as $\mu_{i,y_j,e} = \sum_{t_{jk} \in A_{i,e}} o_{ijke}$.

5 Methodology

We present details about the team interaction protocol, agent behavior, evaluation metrics, and experiment design in this section.

5.1 Interaction Protocols

We describe the protocols that govern the human-agent ad hoc teamwork. Two interaction protocols have been designed to guide task allocation process in an ad hoc environment: (i) the Human Allocator Protocol and (ii) the Agent Allocator Protocol. The former assigns the task allocator role to the human teammate, and is illustrated as follows:

1. *The protocol asks agent teammate for its task types confidence levels.*
2. *The protocol passes the agent's confidence levels to the human.*
   The following steps comprise an episode and are repeated N *times*
   Episode starts: $e \leftarrow 1$

3. *The protocol asks Human to provide task allocations for the team.*
4. *Allocated tasks are assigned to the team members.*
5. *The protocol receives human and agent task performance measures and computes statistics.*
6. *The protocol displays team overall team performance as well as individual team member performances for the episode on their respective task boards.*
   Episode ends
   $e \leftarrow e + 1$*; if* $(e < N)$*, Go to step 3*

The Agent Allocator Protocol is the flip side of the coin and assigns the task allocator role to the agent. Team members repeatably interact over different stages in both protocols: Task Allocation, Task Completion, and Taskwork results (see Though these protocols provide a framework for team interaction and task allocation, they do not dictate the allocation strategy used by the allocator. For the current study, we use a perfect information scenario, where all team information, such as set of team tasks, task assignments to team members, and the task performance is fully observable for all team members.

5.2 Agent Characteristics

Expertise: We configure an agent team member with a fixed expertise profile that has different expertise level for different task types, represented as a vector of probabilities for successful completion of task types[1]. **Agent Allocator Strategy:** In the current paper, we also use the following additional constraints within the CHATboard framework that informs the allocator strategy. We assume each task is allocated to and performed by a single team member and does not require work from multiple individuals, i.e., $A_{i,e} \cap A_{j,e} = \phi$. We additionally required that the total number of tasks assigned to each team member be the same, i.e., $\forall x, y, |A_x| = |A_y|$. Different number of tasks can however be assigned to two team members for different task types.

The primary allocation goal is to maximize utilization of the available team capacity given the expertise of the team. Additionally, agent should account for the constraint that team members have to do equal number of task items. Instead of using *task items* for task division, the agent uses *task types*. The agent stores and uses estimates of on task completion rates by task types for the human team member in the allocation procedure.

$$Max \sum_{y \in M} (x_y a(y) + (1 - x_y)h(y)); s.t. \forall y, x_y \in 0, 1$$

$$\sum_{y \in M} x_y = \sum_{y \in M} (1 - x_y) = \frac{|M|}{2}.$$

[1] Agent expertise is simulated in our experiments: given a expertise (confidence) level P_t of the agent for task type t, a task of type t is considered successfully completed if a coin flipped with probability P_t returns head; else failure is reported on the task.

Algorithm 1 Agent Allocator Strategy

Input: N= $\{p_h, p_g\}$, M= $\{y_1,\ldots,y_m\}$, E

1: **for** $e = 1....E$ **do**
2: **if** $e = 1$ **then**
3: $Q_{i,y_j} \leftarrow p_i(y_j)$, $\forall p_i \in N, y_j \in M$
4: each T_{y_j} is partitioned into n equal size subsets, which are randomly allocated to agent i to form $A_{i,1}$, for each $p_i \in N$
5: **else**
6: $A_{i,e} \leftarrow$ `getAllocations`$(Q_{i,e})$
7: **end if**
8: **if** y_j is allocated to p_i **then**
9: $Q_{i,y_j} \leftarrow (1-\alpha) \cdot Q_{i,y_j} + \alpha \cdot \mu_{i,y_j,e}$
10: **end if**
11: **end for**

In the above equations, x_y is binary variable indicating whether a task type, y, is assigned to human or agent, based on the current performance estimate of the human, $h(y)$, and agent, $a(y)$, on that task type. As per requirement, each team member is assigned exactly half of the task types. This is an *unbalanced assignment problem*, as number of task types is greater than number of team members ($m > n$). It can be solved by transforming it into a *balanced* formulation, e.g., adding dummy variables, and running, e.g., Hungarian algorithm [19]. We utilize the SCIP mixed integer programming solver [29], represented by `getAllocations()` procedure in Line 6 of Algorithm 1, to find the allocation that maximizes utilization of team's confidence levels.

In many task allocation formulations, e.g., matching markets, assignment problems, and others, participants' preferences or confidence levels are assumed to be accurately known [35]. In our formulation, however, learning is needed as we believe human participant's estimates of their capabilities can be inaccurate. The second goal that agent's strategy should account for is related to learning and adaptation. Since this is an ad hoc environment, the second goal of our agent is to quickly learn about its partner's expertise levels and quickly adapt the allocations accordingly for improved team performance. After each interaction, e, the agent updates the capability model, Q_{i,y_j}, of team member, p_i, for each task type, y_j, from the observed performances, $\mu_{i,y_j,e}$, as follows: $Q_{i,y_j} \leftarrow (1-\alpha) \cdot Q_{i,y_j} + \alpha \cdot \mu_{i,y_j,e}$. In the first episode, however, the agent allocator explores team member's capabilities by partitioning task items within each task type, T_{y_j}, equally among team members, as shown in Line 4 in Algorithm 1.

5.3 Evaluation Metrics

Human Teammate Miscalibration and Variability Trends: In our experiments, human teammates coordinate with agent to accomplish tasks items from m task types (we have used $m = 4$ in our experiments). We measure the variability, over task types, of the difference between the human teammates' stated confidence levels and their actual performance.

The confidence levels shared by a human teammate for each task type are used as estimated probability of success for the respective task types. The agent maintains a moving average over the episodes of the team member's performance on a task type as the percentage of tasks of that type that the human successfully completes. We measure miscalibration for a human player i for task type y_j, based on the stated confidence level, $p_i(y_j)$, and actual average performance on that task type over all episodes, $\mu_{i,y_j} = \frac{1}{E}\sum_{e=1}^{E}\mu_{i,y_j,e}$, as squared error: $Miscalibration_{i,y_j} \leftarrow (p_i(y_j) - \mu_{i,y_j})^2$.

Team Performance: Human and agent collaborate as a team to complete the set of tasks. We consider boolean task completion: a task allocated to a team member is either successfully completed or a failure is reported. Team overall performance is measured as the percentage of successful completion of assigned tasks over all episodes: Unweighted Team Performance is measured as the average team performance over episodes, $\frac{1}{E}\sum_{e=1}^{E}R_{team,e}$, where $R_{team,e}$ is the team performance in episode e, which is the average performance, μ, of all team members over all task types in that episode $R_{team,e} \leftarrow \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{i,y_j,e}$.
Team Improvement and Learning: Since our scenario is ad hoc, it requires quick learning and improvements in team performance from task allocators. We investigate the differences in mean performance between episodes to gauge improvements. We also measure the ability to improve as the weighted team performance over episodes, with the performance of latter episodes are weighted more than the earlier ones: Weighted Team Performance $\leftarrow \frac{1}{E}\sum_{e=1}^{E}z_e \cdot R_{team,e}$, where z_e is the weight for episode e.
Potential Realization: An effective allocator will better utilize the capacity of the team and realize as much of their teammate's potential as possible. Potential realization can be measured through the difference between available capacity and utilized capacity. We have perfect knowledge of the agent's capacity, which is fixed at design time. We do not know, however, know of the available capacity of human team members. We compare the difference in the capacity utilized by human and agent allocators. We measure utilized capacity of humans as the individual performance level within the team. The performance (success rate) of an agent i over all episodes, referred to as *Potential Realization* of i, is $S_i = \sum_{e=1}^{E}\sum_{y_j \in M}\mu_{i,y_j,e}$. We designate by S_i^h and S_i^a the performance (potential realization) of agent i under human and agent allocator protocol respectively.
Weighted Likeability: The human-agent team is expected to accomplish m task types over the interaction episodes. At the end of the study, we ask human participants how much they liked each task type by asking them to rate their likeability of each task type on a 10-point Likert scale. For each participant, p_i, we compute the weighted likeability over all allocated tasks as $\sum_{y_j \in M} l_{i,y_j}\sum_{e=1}^{E}|A_{i,y_j,e}|$, where $A_{i,y_j,e}$ is the set of tasks of type y_j allocated to player p_i in episode e and l_{i,y_j} is the human player p_i's stated likeability of task type y_j.

5.4 Experimental configurations

We conduct experiments with teams of one human and one agent ($n = 2$), i.e., $N = \{p_a, p_h\}$. We use four task types ($m = 4$), i.e., M: $\{y_1, y_2, y_4, y_4\}$, which are *Identify Language*, *Solve WordGrid*, *Identify Landmark*, and *Identify Event* (examples of these task types shown in Figure 2). The task types in this paper are selected so that, for each type, sufficient expertise variations in recruited human subjects are likely. For example, *Identify Language* is a task type in which team are asked to identify the language, e.g. Japanese, in a text message from a number of options, e.g., Japanese, German, Hebrew, Arabic.

We created 32 ($r = 8$) task item instances for each episode, and the total number of interactions is four, $E = 4$. The confidence levels are stated in a [1,100] range, which are then scaled by the agent internally into a [0,1] to be interpreted as probabilities of completing tasks of that type. Also, we configure the agent strategy with $\alpha = 0.4$ since Ad hoc situations require allocation strategies to quickly learn about team's capabilities. Additionally, for the weighted performance measure, we have used the following vector of weights over episodes: $z = [0.15, 0.20, 0.30, 0.35]$; it assign more value to performance on latter episodes (any weights that does that would qualitatively produce similar results).

We recruited 130 participants from Amazon Mechanical Turk, 65 for each condition, as is recommended for a medium-sized effect [7]. We use a between-subject, and each team is assigned randomly to one protocol or the other. After participants agree to the Informed Consent Form, they read a description of the study, and then start the first episode. Each episode contains three phases: taskwork allocation, taskwork completion, and taskwork results. After each episode, the results are displayed to both human and agent teammates, which include overall and per-type performance levels. Once participants complete all four episodes, they are asked to complete a survey including their satisfaction on various aspects of teamwork and their likeability for task types. We incorporate random comprehension attention checks to ensure result fidelity [15]. Participants receive a bonus payment based on team performance.

6 Experimental Results

We now summarize the principal experimental results.

Human Variability and Miscalibration: We analyze human variability and task type perceptions in their stated confidence levels and their performance. We first analyze human variability in their stated confidence levels using one-way ANOVA. We find that confidence level between task types ($M_A = 63.27, SD_A = 23.16, M_B = 57.01, SD_B = 21.45, M_C4 = 77.64, SD_C = 19.06, M_D = 41.49, SD_D = 21.70$) are significantly different, F=31, $p < 0.001$. We similarity evaluate variability in humans' actual performances and find that actual performance levels between task types ($M_A = 77.52, SD_A = 17.01, M_B = 75.87, SD_B = 16.28.45, M_C4 =$

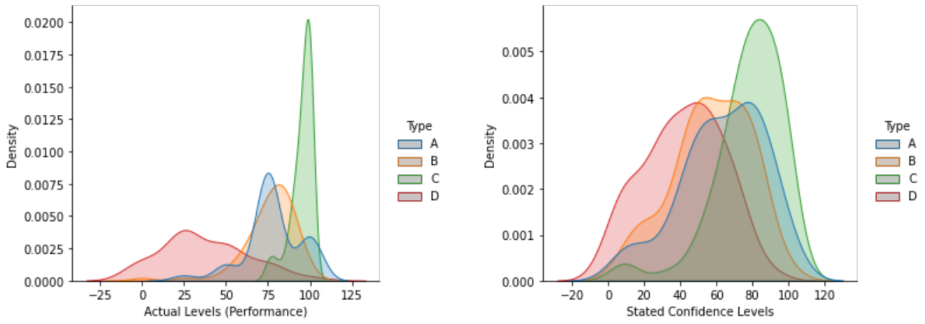

Fig. 3. Human Variability in Stated Confidence (Right) and Actual Performance (Left).

Table 1. Stated Levels and Performances for task types.

Task Type \ Level	Stated		Actual	
	Mean	SD	Mean	SD
Identify Language (**A**)	63.27	23.16	77.52	17.01
Identify Landmark (**B**)	57.01	21.45	75.87	16.28
Solve WordsGrid (**C**)	77.64	19.06	95.0	6.4
Identify Event (**D**)	41.49	21.70	37.30	25.43

95.0, $SD_C = 6.4$, $M_D = 37.30$, $SD_D = 25.43$) are significantly different, F=123, p < 0.001. As Figure 3 and Table 1 show, humans are exhibiting variability and different perceptions toward the task types. **H0a** *is supported.*

We analyze confidence levels estimates stated by human teammates in the Agent Allocator Protocol for the different task types: A, B, C, and D. We analyze the average squared error of the difference between the stated confidence level and actual performance over all task types, 0.08, and was found to be significantly different from zero, t = 7.4, p < 0.001. We then compute the squared error for each task type ($M_A = 0.07$, $SD_A = 0.13$, $M_B = 0.08$, $SD_B = 0.13$, $M_C = 0.06$, $SD_C = 0.12$, $M_D = 0.12$, $SD_D = 0.14$), and find that it is significantly different from zero, $t_A = 4.37$, $p_A < 0.001$, $t_B = 5.28$, $p_B < 0.001$, $t_C = 4.16$, $p_C < 0.001$, $t_D = 7.11$, $p_D < 0.001$ (See Figure 4). Thus, human teammates are showing miscalibration tendencies in all task types. **H0b** *is supported.*

To determine whether human teammates are over- or under-estimating their stated confidence levels in different task types, relative to actual performance, we run non-parametric Sign Tests. We found that, on average, human tend to under-estimate their capabilities relative to actual performance (S_{avg}=18,p_{avg}=0.001). We then run Sign Test for each task type, and find that human teammates are significantly underestimating their capabilities for task type A, B, and C (S_A=15,p_A<0.001, S_B=13,p_B<0.001, S_C=7,p_C<0.001), and over-estimating for task type D (S_D=38,p_D=0.018).

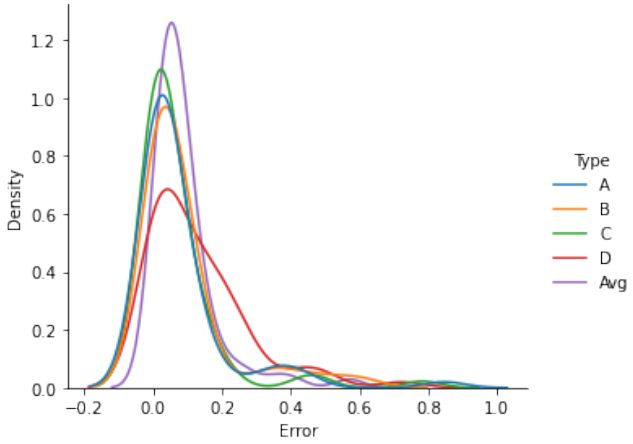

Fig. 4. Density of Squared Estimation Error for task types.

We analyze task type characteristics, and found that task type A, B, and C share one common trait in which they are more general and familiar to typical human teammates, whereas task type D, *Identify Event*, is more specialized [2].

Team Performance: The teams using Agent Allocator Protocol (M = 0.75, SD = 0.04) compared to ones using Human Allocator Protocol (M = 0.69, SD = 0.09) demonstrated significantly higher team performance, t = 4.4, p $<$ 0.001, with a large size effect, cohen's d=0.86 (See Table 2). **H1** *is supported.*

Learning And Improvement: Since the teams are working in an ad hoc environment, task allocators need to quickly learn about team capabilities and increase team performance. First, we investigate if team performances over episodes is different in each protocol. We find that it is significantly different for the Agent Allocator Protocol ($M_{eps1} = 0.59, SD_{eps1} = 0.10, M_{eps2} = 0.76, SD_{eps2} = 0.11, M_{eps3} = 0.82, SD_{eps3} = 0.10, M_{eps4} = 0.83, SD_{eps4} = 0.11$), $F_a = 167.17$, $p_a < 0.001$. We also find that it is significantly different for the Human Allocator Protocol ($M_{eps1} = 0.66, SD_{eps1} = 0.10, M_{eps2} = 0.67, SD_{eps2} = 0.13, M_{eps3} = 0.71, SD_{eps3} = 0.12$, $M_{eps4} = 0.71, SD_{eps4} = 0.12$), $F_h = 3.17$, and $p_h = 0.024$.

The agent allocator starts has lower performance ,$M_{eps1} = 0.59$, than human allocator, $M_{eps1} = 0.66$ in the first episode. This is due to the agent strategy of exploration during the first episode. However, the agent improves quickly, and outperforms human in the second, third, and fourth episodes. The agent improves team performance by a significant margin going from episode 1 to episode 2, and then by smaller margins going from episode 2 to episode 3, and episode 3 to episode 4. The improvements over episodes by the Human allocator is less pronounced.

Moreover, we run Post hoc analysis, using Tukey's HSD Test, to evaluate the performance differences between episodes (See Figure 5). When Human is allocating, we find no significant mean differences between the episodes, $E2 - E1 = 0.007, p = 0.98, E3 - E1 = 0.05, p = 0.10, E4 - E1 = 0.05, p = 0.08, E3 -$

Table 2. Team Performance (*p $<$ 0.001).

Performance \ Allocator	Human		Agent		
	Mean	SD	Mean	SD	t
Unweighted	0.69	0.09	**0.75**	0.04	4.4*
Weighted	0.70	0.10	**0.78**	0.04	5.8*

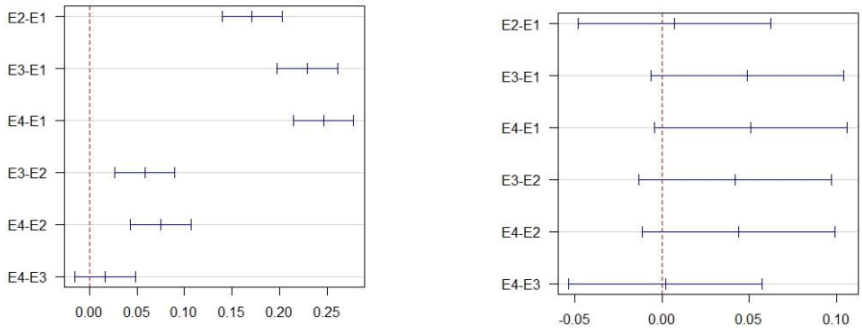

Fig. 5. Tukey's HSD Test: Differences in mean levels of four episodes (E1 to E4). Left: Agent, Right: Human.

$E2 = 0.04, p = 0.20, E4 - E2 = 0.42, p = 0.17, E4 - E3 = 0.001, p = 0.99$. We do, however, find significant mean differences between episodes with the Agent Allocator, except for E4-E3, $E2 - E1 = 0.17, p < 0.001, E3 - E1 = 023, p < 0.001, E4 - E1 = 0.25, p < 0.001, E3 - E2 = 0.06, p < 0.001, E4 - E2 = 0.08, p < 0.001, E4 - E3 = 0.02, p = 0.52$. This shows that the agent is, indeed, improving after each experience. One possible interpretation between the small difference between episode 3 and 4, relative to the larger differences from episodes E1 to E2, and from E2 to E3, is that the agent is getting close to the optimal allocation of tasks based on the team member capabilities.

We also note that performance of teams using the Agent Allocator Protocol (M = 0.78, SD = 0.04) are better than teams using the Human Allocator Protocol (M = 0.70, SD = 0.10) in weighted performance, t = 5.8, p $<$ 0.001. In other words, the agent is showing better learning of its teammate's capabilities and adapting the task allocations accordingly to further improve team performance in latter rounds. since weighted performance measures overall team performance over the latter, rather than, earlier episodes. The agent allocator significantly outperforms the human allocator using the weighted performance measures (See Table 2). **H2** *is supported.*

Potential Realization: We compared teams based on how allocators realize potential of teammates and themselves. The pertinent question is: which alloca-

Table 3. Self, teammate potential Realization by allocators.

Performance \ Allocator	Human		Agent	
	Mean	SD	Mean	SD
Human	0.81	0.10	**0.87**	0.06
Agent	0.59	0.12	**0.74**	0.05

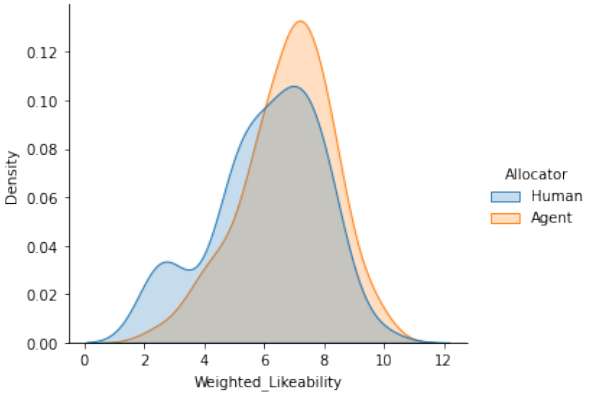

Fig. 6. Weighted Likeability Density for Human and Agent Protocols.

tor utilizes human capacity better? We find that teams who have agents as task allocators (M = 0.87, SD = 0.06) realize significantly more human potential than Human Allocator (M = 0.81, SD = 0.10), t = 2.2, p = 0.02. **H3** *is supported.*

We also analyze how team allocators effectively utilize agent capacity. We find that agent capacity utilization or performance is significantly higher in teams who have agents as task allocators (M = 0.74, SD = 0.05) compared to teams with Human allocators (M = 0.59, SD = 0.12), t = 5.02, p $<$ 001. Thirdly, we investigate which allocator utilizes the capacity of their teammate better. We find that teams who Agent allocators (M = 0.87, SD = 0.06) significantly realize more performance from their teammates than Human Allocator (M = 0.59, SD = 0.12), t = 13.4, p $<$ 0.001.

We do not analyze self-realization between human and agent allocators since human capacity in the Human Allocator Protocol is unknown. We also define the level of agent capacity or confidence level structure prior to the interaction; thus, we cannot compare self-realization of human and agent allocators. We posit, however, when allocators are agents, they realize more potential in the team; both in themselves and in the human team member (See Table 3) [2].

Weighted Likeability: To understand the performance differences between the Human and Agent Allocator Protocols, we analyze the task types allocated to

[2] Humans outperform agents for both allocators as agents are endowed with medium-level capabilities. Increasing agent expertise will change relative performances.

human teammates. Do humans allocate more tasks of types they like to themselves? We find that Agent allocators ($M_a = 6.77, SD_a = 1.51$) allocate more items of liked task types to the human team member than does the human allocator ($M_h = 6.07, SD_h = 1.80$), $t_{like} = 2.3, p_{like} = 0.01$ (See Figure 6).

7 Conclusions and Future Work

We introduced CHATboard, a flexible task allocation framework between human and agent team members for ad hoc scenarios. While CHATboard can be configured to support larger teams and more complex constraints between tasks, such as multiple workers per task, in this paper we showed its efficacy in supporting coordination between one human and one autonomous agent.

To understand team dynamics with respect to task allocation within human-agent teams, we presented two interaction protocols and team designs in which task allocator role is either assigned to human or agent team member: Human and Agent Allocator Protocols. We ran experiments with these team designs and showed human teammates often exhibit miscalibration, where they either over- or under-estimate their capabilities.

We demonstrated that agent task allocators generally increase the quality of team with respect to team performance and realizing potential of team compared to human allocators. The agent allocators learn quickly about team capabilities, and realize more potential in the team, both their own and of their human teammate. Our analysis of the experiments also confirms various hypotheses we had posed about such ad hoc human-agent team coordination.

Though finding the reason for the lower performance of human allocators is beyond scope of this paper, we conjectured that it might be due to humans allocating more tasks they like to themselves, even though they may not be good at it. We find, however, that the agent is allocating more likeable tasks to the human teammate. The lower performance might be explained by biases identified in behavioral economics, such as prospect theory, in which they perceive performance gains or success differently than losses or failure rate. We leave this line of investigation to future work.

As future work, we plan to work on better understanding the performance of humans as allocators, e.g., what explains the lower performance of human-agent team with Human allocators. We will evaluate the effect of different agent expertise distributions on team performances. We also plan to experiment with different environment configuration, including those where the constraint of equal division of tasks is relaxed. Another future research direction is making the agent strategy more robust to human miscalibration tendencies. Lastly, we plan to study how the dynamics of human-agent teams change when the team consists of more than two members. Having better grasp of these directions can inform our human-agent team design with respect to task allocation

References

1. Abdulrahman, A., Richards, D., Bilgin, A.A.: Reason explanation for encouraging behaviour change intention. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. pp. 68–77 (2021)
2. Adams, P.A., Adams, J.K.: Confidence in the recognition and reproduction of words difficult to spell. The American journal of psychology pp. 544–552 (1960)
3. Anderson, A., Kleinberg, J., Mullainathan, S.: Assessing human error against a benchmark of perfection. ACM Transactions on Knowledge Discovery from Data (TKDD) **11**(4), 1–25 (2017)
4. Athey, S.C., Bryan, K.A., Gans, J.S.: The allocation of decision authority to human and artificial intelligence. In: AEA Papers and Proc. vol. 110, pp. 80–84 (2020)
5. Barrett, S., Stone, P., Kraus, S.: Empirical evaluation of ad hoc teamwork in the pursuit domain. In: AAMAS. pp. 567–574 (2011)
6. Binetti, G., Naso, D., Turchiano, B.: Decentralized task allocation for surveillance systems with critical tasks. Robotics and Autonomous Systems **61**(12), 1653–1664 (2013)
7. Brinkman, W.P.: Design of a questionnaire instrument. In: Handbook of mobile technology research methods, pp. 31–57. Nova Publishers (2009)
8. Dias, M.B., Zlot, R., Kalra, N., Stentz, A.: Market-based multirobot coordination: A survey and analysis. Proceedings of the IEEE **94**(7), 1257–1270 (2006)
9. Ernst, A., Jiang, H., Krishnamoorthy, M.: Exact solutions to task allocation problems. Management science **52**(10), 1634–1646 (2006)
10. Fisher, D.M.: Distinguishing between taskwork and teamwork planning in teams: Relations with coordination and interpersonal processes. Journal of applied Psychology **99**(3), 423 (2014)
11. Genter, K., Agmon, N., Stone, P.: Role-based ad hoc teamwork. In: Proceedings of the Plan, Activity, and Intent Recognition Workshop at the Twenty-Fifth Conference on Artificial Intelligence (PAIR-11) (August)
12. Gerkey, B.P., Matarić, M.J.: A formal analysis and taxonomy of task allocation in multi-robot systems. The International journal of robotics research **23**(9), 939–954 (2004)
13. Gervits, F., Thurston, D., Thielstrom, R., Fong, T., Pham, Q., Scheutz, M.: Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In: AAMAS. pp. 429–437 (2020)
14. Gunn, T., Anderson, J.: Effective task allocation for evolving multi-robot teams in dangerous environments. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). vol. 2, pp. 231–238. IEEE (2013)
15. Hauser, D., Paolacci, G., Chandler, J.: Common concerns with mturk as a participant pool: Evidence and solutions. (2019)
16. Hayes-Roth, B.: A blackboard architecture for control. Artificial intelligence **26**(3), 251–321 (1985)
17. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
18. Korsah, G.A., Stentz, A., Dias, M.B.: A comprehensive taxonomy for multi-robot task allocation. The Intl Journal of Robotics Research **32**(12), 1495–1512 (2013)
19. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
20. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 29–38 (2019)

21. Larson, L., DeChurch, L.A.: Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. The Leadership Quarterly **31**(1), 101377 (2020)
22. Lin, R., Gal, Y., Kraus, S., Mazliah, Y.: Training with automated agents improves peoples behavior in negotiation and coordination tasks. Decision Support Systems (DSS) **60**(1–9) (April 2014)
23. Lott, C., McAuliffe, A., Kuttal, S.K.: Remote pair collaborations of cs students: Leaving women behind? In: 2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). pp. 1–11. IEEE (2021)
24. Mathieu, J.E., Hollenbeck, J.R., van Knippenberg, D., Ilgen, D.R.: A century of work teams in the journal of applied psychology. Journal of applied psychology **102**(3), 452 (2017)
25. Mathieu, J.E., Rapp, T.L.: Laying the foundation for successful team performance trajectories: The roles of team charters and performance strategies. Journal of Applied Psychology **94**(1), 90 (2009)
26. Mosteo, A.R., Montano, L.: A survey of multi-robot task allocation. Instituto de Investigacin en Ingenierła de Aragn (I3A), Tech. Rep (2010)
27. Nunes, E., Manner, M., Mitiche, H., Gini, M.: A taxonomy for task allocation problems with temporal and ordering constraints. Robotics and Autonomous Systems **90**, 55–70 (2017)
28. Patel, R., Rudnick-Cohen, E., Azarm, S., Otte, M., Xu, H., Herrmann, J.W.: Decentralized task allocation in multi-agent systems using a decentralized genetic algorithm. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 3770–3776. IEEE (2020)
29. Perron, L., Furnon, V.: Or-tools, https://developers.google.com/optimization/
30. Puranam, P., Alexy, O., Reitzig, M.: What's "new" about new forms of organizing? Academy of Management Review **39**(2), 162–180 (2014)
31. Ramchurn, S.D., Huynh, T.D., Ikuno, Y., Flann, J., Wu, F., Moreau, L., Jennings, N.R., Fischer, J.E., Jiang, W., Rodden, T., Simpson, E., Reece, S., Roberts, S.J.: Hac-er: A disaster response system based on human-agent collectives. In: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems. pp. 533–541. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2015)
32. Robinette, P., Wagner, A.R., Howard, A.M.: Building and maintaining trust between humans and guidance robots in an emergency. In: AAAI Spring Symposium: Trust and Autonomous Systems. pp. 78–83. Stanford, CA (March 2013)
33. Rosenfeld, A., Agmon, N., Maksimov, O., Kraus, S.: Intelligent agent supporting human–multi-robot team collaboration. Artificial Intelligence **252**, 211–231 (2017)
34. Sanchez, R.P., Bartel, C.M., Brown, E., DeRosier, M.: The acceptability and efficacy of an intelligent social tutoring system. Computers & Education **78**, 321–332 (2014)
35. Shoham, Y., Leyton-Brown, K.: Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press (2008)
36. Zhao, W., Meng, Q., Chung, P.W.: A heuristic distributed task allocation method for multivehicle multitask problems and its application to search and rescue scenario. IEEE transactions on cybernetics **46**(4), 902–915 (2015)