Effective Task Allocation in Ad Hoc Human-Agent Teams

Sami ABUHAIMED and Sandip SEN
Tandy School of Computer Science, The University of Tulsa

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 collaboration 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 collaboration between diverse groups. We study 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 collaboration protocols that differ in the team member responsible for allocating tasks to the team. We designed, implemented, and experimented with an environment in which human-agent teams repeatedly collaborate to complete heterogeneous task sets.

Keywords. Human-agent collaboration, Team performance, Task allocation

1. Introduction

Agents can collaborate with people in critical tasks, including guiding emergency evacuations [22] and disaster relief [21]. Recent intelligent agent applications assume traditionally human roles in human-agent teams, e.g., tutor [24] and trainer [16]. Since human-agent teams are being recognized as a routine and functionally critical important component of our societies, researchers are studying the interactions and dynamics within these teams to understand and improve on their design [8]. Such human-agent teams have been studied in physical (robotic) and virtual settings [23].

We are interested in human-agent collaboration in *ad hoc teams* where team members have no prior knowledge of or interaction experience with their teammate: *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* [7].

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 [11]. 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 has been studied extensively in agent teams [18] as well as in human team and organizations literature [20]. 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 collaboration framework for task allocation and performance analysis: the Collaborative Human-Agent Taskboard (CHATboard). We use CHATboard for ad hoc human-agent team collaboration, 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 collaboration using the Human and Agent Allocation protocols. We present the results and our analysis to confirm our hypotheses and identify interesting phenomena that suggest future research tasks.

2. Related Work

Human-agent teams have been studied in different domains such as space robotics [8], therapy [1], and decision-making [3]. Furthermore, much of the focus has been on agents who play supportive roles to human teammates [14], and they have been studied in robotic and simulation settings [23].

We focus on an ad hoc environment, whereas studies, such as [8], incorporate training or interactions with 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” [15].

Task allocation has been studied extensively in multi-agent teams [12,18]. In agent teams, the focus is on designing efficient mechanisms for agents to distribute tasks within their society. Task allocation is also studied in humans' team and organization literature. The mechanism of task allocation, which includes capabilities identification, role spec-

ification, and task planning, is considered an important component of teamwork [17]. Any organization needs to solve four universal problems, including task allocation, to achieve its goals [20]. 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.

There is a recent focus in agent teams on ad hoc environments [5] in which agents collaborate without pre-collaboration, and most work is focused on simulation and robotic environments. Few researchers have studied task allocation in ad hoc human-agent teams. Moreover, the majority of work is not focused on environments that include human teammates; including humans in agent teams may require new approaches, as we do not know whether the same mechanisms would produce similar results. Thus, the study of task allocation with combined human and agent team members is promising and little work is focused on it [4,23]. 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.

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 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

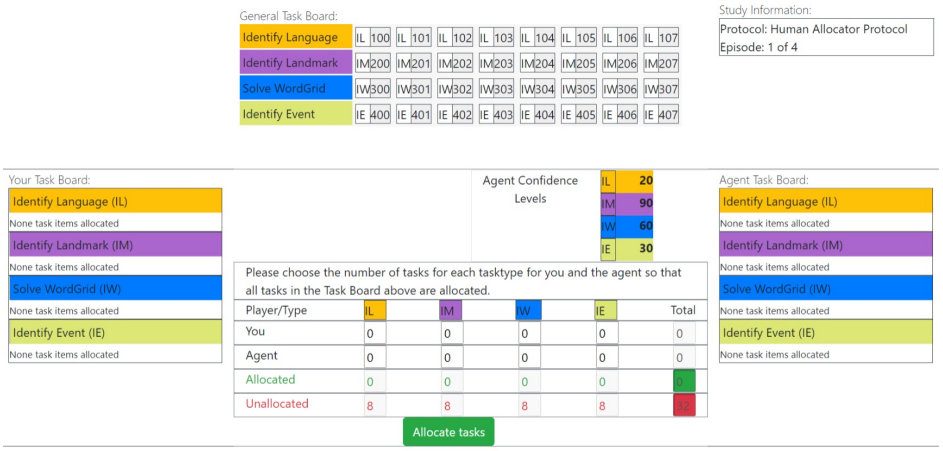

Figure 1. CHATboard showing allocation phase of Human Allocation Protocol.

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.*

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 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 based on the research questions investigated.

We developed CHATboard, an environment that facilitates human-agent, as well as human-human, team collaboration. CHATboard contains a graphical interface that supports human-agent team collaboration to complete a set of tasks (see Figure 1). CHAT-

board 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 collaboration 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 between agents [10]. We incorporate three task boards in our task sharing frame: one shared board, 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. These task boards facilitate collaboration, 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 are designed to guide task allocation process in an ad hoc environment: (i) Human Allocator Protocol and (ii) Agent Allocator Protocol.The former assigns the task allocator role to the human teammate, and works 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. 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: An agent has a fixed profile with different expertise levels for different tasks, represented as a vector of probabilities for successful completion of task types[1].
Agent 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.

$$\max \sum_{y \in M} (x_y a(y) + (1 - x_y) h(y)); s.t. \forall y, x_y \in 0, 1, \text{where } \sum_{y \in M} x_y = \sum_{y \in M} (1 - x_y) = \frac{|M|}{2}.$$

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 [13]. We utilize the SCIP mixed integer programming solver [19], 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 [25]. 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.

[1] Agent expertise is simulated by flipping a coin with success probability of P_t, the confidence level.

Algorithm 1 Agent Allocator Strategy

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

for $e = 1....E$ **do**
  if $e = 1$ **then**
    $Q_{i,y_j} \leftarrow p_i(y_j), \forall p_i \in N, y_j \in M$
    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 $p_i \in N$
  else
    $A_{i,e} \leftarrow$ `getAllocations`$(Q_{i,e})$
  end if
  if y_j is allocated to p_i **then**
    $Q_{i,y_j} \leftarrow (1-\alpha) \cdot Q_{i,y_j} + \alpha \cdot \mu_{i,y_j,e}$
  end if
end for

5.3. Evaluation Metrics

Human Teammate Miscalibration and Variability Trends: In our experiments, human and agent teammates collaborate to accomplish tasks from m task types. 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. The agent maintains a moving average over the episodes of the team member's performance on a task type. 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: 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, 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 realiza-

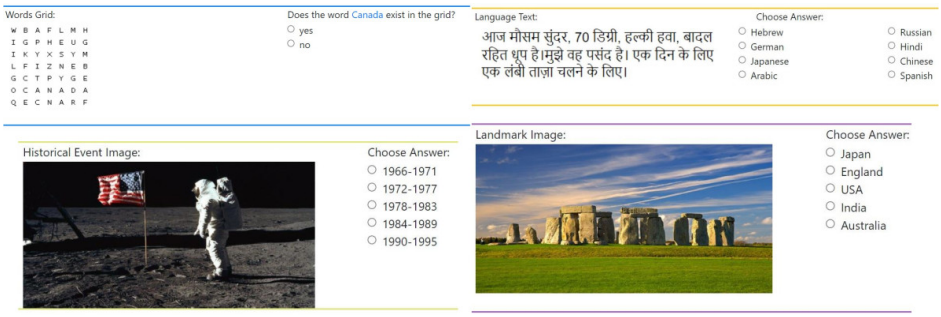

Figure 2. Instances of different task types.

tion) of agent i under human and agent allocator protocol respectively.

Weighted Likeability: After 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 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 ($n = 2$), $N = \{p_a, p_h\}$. We use four task types ($m = 4$), M: $\{y_1, y_2, y_4, y_4\}$, which are *Identify Language*, *Solve WordGrid*, *Identify Landmark*, and *Identify Event* (examples of task types shown in Figure 2). The task types 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 members 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 total number of interactions is four, $E = 4$. The confidence levels are stated in a [1,100] range, which are then scaled by 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 Turk, 65 for each condition, as is recommended for a medium-sized effect [6]. We use a between-subject experimental design, and each team is assigned randomly to a protocol. After participants agree to the Informed Consent Form, they read the study description, and start the first episode. Each episode contains three phases: taskwork allocation, taskwork completion, and taskwork results. After each episode, results are displayed to both human and agent teammates, which include overall and per-type performance levels. Once participants complete four episodes, they are asked about their likeability for task types. We incorporate random comprehension attention checks to ensure result fidelity [9]. Participants receive a bonus payment based on team performance.

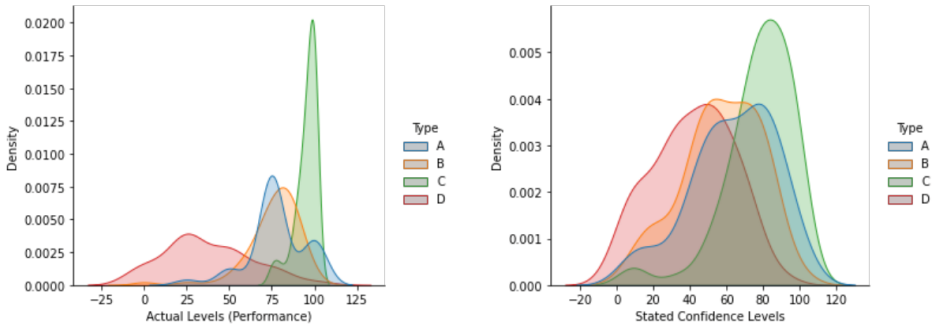

Figure 3. Human Variability in Stated Confidence (Right) and Actual Performance (Left).

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

6. 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 ($M_A = 77.52, SD_A = 17.01, M_B = 75.87, SD_B = 16.28.45, M_C4 = 95.0, SD_C = 6.4, M_D = 37.30, SD_D = 25.43$) 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 if human teammates err in estimating their stated confidence levels in different task types, relative to actual performance, we run non-parametric Sign Tests.

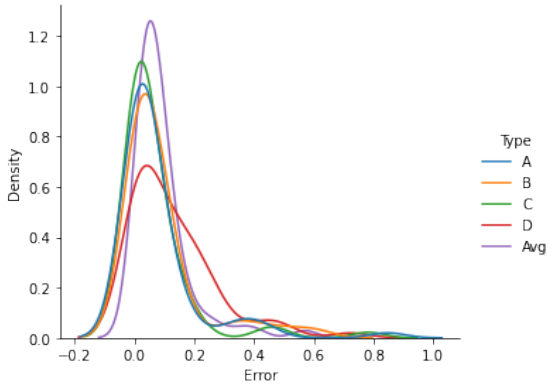

Figure 4. Density of Squared Estimation Error for task types.

We found that, on average, human tend to underestimate 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,$p_A4 < 0.001$, S_B=13,$p_B < 0.001$, S_C=7,$p_C < 0.001$), and overestimating for task type D (S_D=38,$p_D = 0.018$). 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: Given the ad hoc environment, task allocators need to quickly learn about team capabilities and increase team performance. Team performances over episodes is significantly different for both the Agent Allocator Protocol ($M_{eps1} = 0.59, SD_{eps1} = 0.10, M_{eps2} = 0.76, SD_{eps2} = 0.11, M_{eps3} = 0.82, SD_{eps3} = 0.10, M_{eps4} = 0.83, SD_{eps4} = 0.11$), $F_a = 167.17$, $p_a < 0.001$ and 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 ($M_{eps1} = 0.59$) performs worse than human allocatorn ($M_{eps1} = 0.66$) in the first episode due to its initial exploration strategy. 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 between the following episodes. 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. When Human is allocating, we find no significant mean differences between the episodes, $E2 - E1 = 0.007, p = 0.98, E3 - E1 = 0.05, p = 0.10, E4 - E1 = 0.05, p = 0.08, E3 - E2 = 0.04, p = 0.20, E4 - E2 = 0.42, p = 0.17, E4 - E3 = 0.001, p = 0.99$. 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

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*

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 allocator 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 define the level of agent capacity or confidence level structure prior to the interaction. Hence we cannot compare self-realization of human and agent allocators. Agent allocators realize more team potential both in themselves and in the human team member (See Table 3). Humans outperform agents for both allocators as agents are endowed with medium-level capabilities. Increasing agent expertise will change relative performances.

Weighted Likeability: To understand the performance differences between the two allocation protocols, we analyze the task types allocated to teammates. 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 5).

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

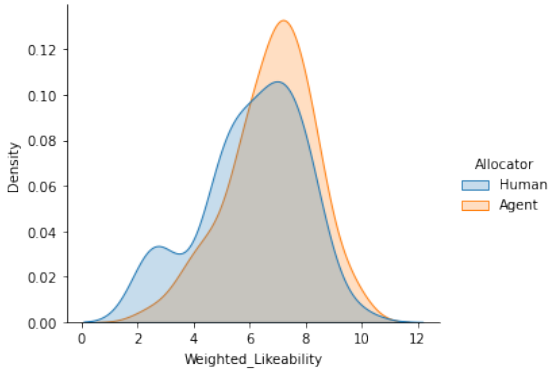

Figure 5. Weighted Likeability Density for Human and Agent Protocols.

7. Conclusions and Future Work

We introduced CHATboard, a flexible task allocation framework between human and agent team members for ad hoc scenarios. We showed its efficacy in supporting collaboration between one human and one autonomous agent. CHATboard can be configured to support larger teams and more complex constraints between tasks.

To understand team dynamics with respect to task allocation within human-agent teams, we investigated interaction protocols and team designs in which task allocator role is either assigned to human or agent team member. We ran experiments with these team designs and showed that 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 collaboration.

We plan to work on better understanding the performance of humans as allocators, e.g., what explains the lower performance of teams with Human allocators. We will evaluate the effect of different agent expertise distributions on team performances. We plan to experiment with different environment configuration, including those where the constraint of equal division of tasks is relaxed. Lastly, we we plan to study how the dynamics of human-agent teams change when the team consists of more than two members.

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] Brinkman, W.P.: Design of a questionnaire instrument. In: Handbook of mobile technology research methods, pp. 31–57. Nova Publishers (2009)

[7] 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)

[8] 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)

[9] Hauser, D., Paolacci, G., Chandler, J.: Common concerns with mturk as a participant pool: Evidence and solutions. (2019)

[10] Hayes-Roth, B.: A blackboard architecture for control. Artificial intelligence **26**(3), 251–321 (1985)

[11] Kahneman, D.: Thinking, fast and slow. Macmillan (2011)

[12] 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)

[13] Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)

[14] 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)

[15] 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)

[16] 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)

[17] 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)

[18] Mosteo, A.R., Montano, L.: A survey of multi-robot task allocation. Instituto de Investigacin en Ingenierła de Aragn (I3A), Tech. Rep (2010)

[19] Perron, L., Furnon, V.: Or-tools, `https://developers.google.com/optimization/`

[20] Puranam, P., Alexy, O., Reitzig, M.: What's "new" about new forms of organizing? Academy of Management Review **39**(2), 162–180 (2014)

[21] 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)

[22] 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)

[23] Rosenfeld, A., Agmon, N., Maksimov, O., Kraus, S.: Intelligent agent supporting human–multi-robot team collaboration. Artificial Intelligence **252**, 211–231 (2017)

[24] 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)

[25] Shoham, Y., Leyton-Brown, K.: Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press (2008)