# Relative Effects of Positive and Negative Explanations on Satisfaction and Performance in Human-Agent Teams

**Bryan Lavender***, **Sami Abuhaimed***, **Sandip Sen**
The University of Tulsa
bal7708, saa8061, sandip@utulsa.edu

## Abstract

Improving agent capabilities and increasing availability of computing platforms and Internet connectivity allows for more effective and diverse collaboration between human users and automated agents. To increase the viability and effectiveness of human-agent collaborative teams, there is a pressing need for research enabling such teams to maximally leverage relative strengths of human and automated reasoners. We study virtual and ad-hoc teams, comprising a human and an agent, collaborating over a few episodes where each episode requires them to complete a set of tasks chosen from given task types. Team members are initially unaware of the capabilities of their partners, and the agent, acting as the task allocator, has to adapt the allocation process to maximize team performance. The focus of the current paper is on analyzing how allocation decision explanations can affect both user performance and the human workers' outlook including factors such as motivation and satisfaction. We investigate the effect of explanations provided by the agent allocator to the human on performance and key factors reported by the human teammate on surveys. Survey factors include the effect of explanations on motivation, explanatory power, and understandability, as well as satisfaction with and trust / confidence in the teammate. We evaluated a set of hypotheses on these factors related to positive, negative, and no-explanation scenarios through experiments conducted with MTurk workers.

## Introduction

With human-agent teams gaining acceptance as practical and effective frameworks in modern societies, researchers are improving their design and developing a deeper understanding of the interactions and dynamics within these teams (Gervits et al. 2020). Human-agent teams have been studied in physical (robotic) and virtual settings (Rosenfeld et al. 2017) and critical areas, including guiding emergency evacuations (Robinette, Wagner, and Howard 2013) and disaster relief (Ramchurn et al. 2015).

We are interested in human-agent collaboration in *ad hoc teams* where team members do not have prior knowledge or interaction with their teammates (Genter, Agmon, and Stone ). Collaboration in ad hoc teams is challenging due to a lack of prior knowledge and established relationships. In this paper, we consider ad hoc human-agent teams repeatedly trying to complete a set of tasks chosen from diverse task types. We expect different human users to have different competences and expertise on various types of tasks. We use a fixed agent expertise distribution (simulated) over the task types. To optimize the performance of a given human-agent team, task allocation process between team members must utilize the relative expertise of team members on different types of tasks. The allocation problem is exacerbated by the fact that a team member does not know the expertise levels of its partner *a priori*. Although we allow human and agent partners to share their estimated expertise on different types of tasks, the accuracy and consistency of these estimates expressed by humans are unreliable (Kahneman 2011). The success of such ad hoc human-agent teams in completing assigned team tasks will critically depend on effective adaptability in the task allocation process.

We found, in pilot experiments, that teams where the agent allocated tasks outperformed teams with human allocators. However, humans expressed greater satisfaction with the process when they were allocators. This presents a *dilemma*: If the human team members are dissatisfied, they may not continue with a team even if the team performs better. Some responses to free-form text questions in surveys associated with these experiments suggested that dissatisfaction of human team members may be caused by their lack of understanding of the rationale behind the task allocation process used by the agent allocator.

Explanations can be a key tool for promoting transparency, interpretability, and trust in AI-based systems. Explanations can be effective in improving the understanding of human teammates about task allocation decisions by agent teammates (Barredo Arrieta et al. 2020; Miller 2019). We evaluate different types of explanations to gauge the most effective approach to adopt. The key research question that we investigate in this paper is the relative effects of no explanation, positive and negative explanations, on human team member satisfaction and user performance.

We use a human-agent team collaboration framework for task allocation and performance analysis, the Collaborative Human-Agent Taskboard (CHATboard), for repeated team task allocation scenarios, with human workers recruited from the Amazon Mechanical Turk (MTurk) platform. We present some conjectures as hypotheses about the availability or absence of explanations and the relative effects of


*Equal Contribution

positive or negative contrastive explanations. We ran experiments that involved repeated collaboration with agent task allocators and varying the types of explanation provided to human teammates for allocation decisions. We present the results and our analysis to confirm our hypotheses and identify interesting phenomena that suggest future research tasks.

## Related Work

Task allocation has been extensively studied in agent teams (Mosteo and Montano 2010), where the focus is on designing efficient mechanisms for agents to distribute items or tasks within their society. It is also studied in the literature on human team and organizations (Puranam, Alexy, and Reitzig 2014), where the task allocation mechanism, which includes capability identification, role specification, and task planning, is considered important components of teamwork (Mathieu et al. 2017).

Most of the work on the use of explanations in AI, XAI, is focused on increasing the transparency of black-box machine learning models (Barredo Arrieta et al. 2020). Most of this work assumes off-line computing, which analyzes learned models and highlights aspects such as feature importance and their effects. Our work uses explanations in an online, interactive modality where explanations have to be generated and presented to human teammates in real-time. There is little existing work on the use of explanations in human-agent interaction and teams that seeks to provide transparency and better understanding of the reasoning behind a decision made by an agent (Barredo Arrieta et al. 2020). Trust and confidence in agent decisions have been shown to increase in human collaborators and teammates when explanations leading to better understanding of agent decision factors are presented to teammates (Shin 2021).

We focus on human-agent collaboration and improving satisfaction and performance with explanations (Neerincx et al. 2018). Explanations have been shown to aid human-agent interactions and teamwork and increase the fidelity of human understanding of a model (Mualla et al. 2022).

(Barredo Arrieta et al. 2020) provides the XAI nomenclature and concepts that are useful for developing systems in which participants are selected from the population at large. *Understandability*, or intelligibility, is the characteristic of a model to make a human in any domain understand its functionality. This is important to measure, as understanding the model can affect outcomes in human decisions. *Explanatory power* is the ability for an explanation to provide understanding to a human interpreting the explanation, as mentioned by (Ehsan et al. 2019a).

Confidence, understandability, explanatory power, and satisfaction are all subjective terms that need to be studied empirically. A survey presented by (Miller 2019) argues that explanations are social in nature and can be derived from the social sciences. Miller argues that answering 'why-question' is contrastive and that humans better understand causal links. We design our explanation generation approach based on this principle. *Contrastive explanations* describe "why event A occurred as opposed to some alternative event B" according to (Kim et al. ). Building on this idea,

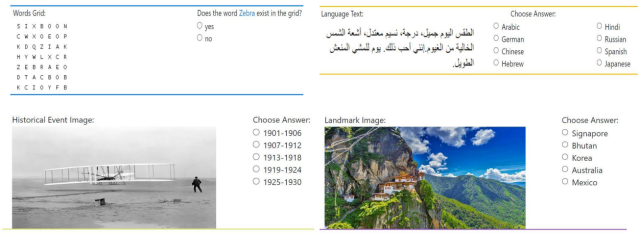


Figure 1: Instances of different task types.

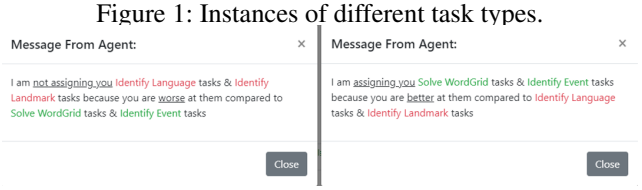


Figure 2: Positive and negative explanations of allocations.

(Mualla et al. 2022) used psychology-based surveys to empirically measure human-subjective concepts such as trust, confidence, satisfaction, or understanding using the Likert scale.

Studies that investigate the role of explainability in task allocation within teams composed of humans and autonomous agents in ad hoc environments are limited. It is important to study the effects of explanations as adequate explanations are subjective to those who interpret them.

## Human-Agent Task Collaboration

We now formally define the repeated human-agent collaboration framework for completing tasks over multiple episodes. 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 $j$. 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 $\mu_{i,y_j,e} = \sum_{t_{jk} \in A_{i,e}} o_{ijke}$.

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 1). 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, Arabic.

## Research Hypotheses

We now motivate and present research hypotheses related to different explanation mechanisms for task allocation in ad

hoc human-agent teams and their relation to different factors including motivation, explanatory power, understandability, confidence, and satisfaction that we will experimentally evaluate in this paper.

Task allocation protocols provide mechanisms for how humans and agents can work together. In the Agent Allocator Protocol, the agent allocates tasks to the team. We observed that agent allocators produce higher team performance compared to human allocators. However, we also observed that human satisfaction is low when agents allocate tasks. The question is how do we increase human satisfaction when an agent allocates? What other human factors are critical to making agent allocators more useful?

We conjectured that providing explanations of agent decisions might improve human perceptions. We analyze agent-specific factors (understandability, confidence and satisfaction in agent) and explanation-associated factors (motivation and explanatory power).

In the initial work, the agent did not provide any rationale for task allocations. We expect this no-explanation condition to have the lowest understandability and human ratings of agent compared to when explanations are provided. When the agent explains the rationale behind the allocation decisions of tasks, the agent has to choose the type of explanation. One of the important criteria humans use to provide explanations (Miller 2019) is that they are contrastive, that is, why we chose X rather than Y. We develop positive and negative contrastive explanations specific to our task allocation context (more details in the Methodology section). Positive explanations highlight the strengths of the participants, whereas negative ones highlight weaknesses. We expect human participants to have higher ratings for positive explanations in all fields. We now list the specific hypotheses to be tested experimentally:

**Hypothesis 1a (H1a)***: Explanations make it easier for the users to understand agent's task allocation decisions.*

**Hypothesis 1b (H1b)***: Positive and Negative Explanations are equally useful for the users to understand the agent's task allocation decisions.*

**Hypothesis 2a (H2a)***: Explanations make users more confident and trusting in the agent.*

**Hypothesis 2b (H2b)***: Positive explanations make users more confident and trusting in the agent compared to negative explanations.*

**Hypothesis 3a (H3a)***: Users will be more satisfied with the agent that provides explanations than the agent who does not provide explanations.*

**Hypothesis 3b (H3b)***: Satisfaction will be higher with agents that provide positive explanations than negative explanations.*

**Hypothesis 4 (H4)***: Positive explanations motivate users more than negative explanations.*

**Hypothesis 5 (H5)***: Positive and negative explanations have equal explanatory power.*

**Hypothesis 6a (H6a)***: User performance will be higher when agent provides explanations.*

**Hypothesis 6b (H6b)***: User performance will be higher when agent provides positive compared to negative explanations.*

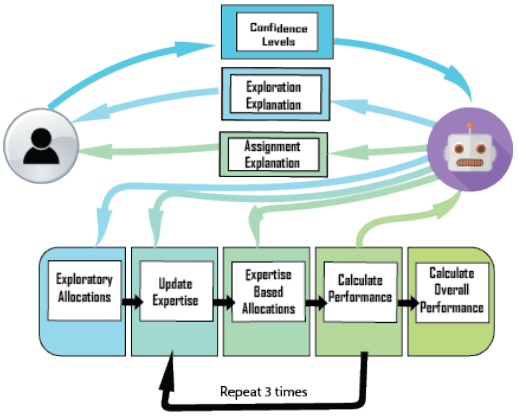


Figure 3: Episode Components.

## Methodology

We now present the team interaction protocol, agent behavior, and evaluation metrics (see Figure 3).

### Interaction Protocol

We describe the protocol that governs the human-agent ad hoc teamwork. In this study, the agent allocates tasks for the human-agent team.

1. *The protocol asks agent for its task types confidence levels.*
2. *The protocol passes the agent's confidence levels to the human. Following steps comprise an episode and are repeated* N *time*
   **Episode starts**: $e \leftarrow 1$
3. *The protocol asks agent 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. *Protocol displays both team and individual team member performances for the episode on respective task boards.*
   **Episode ends**
   $e \leftarrow e + 1$*; if* $(e < N)$*, Go to step 3*

Team members repeatably interact over different stages in the protocol: Task Allocation, Task Completion, and Taskwork results. Though the protocol provides a framework for team interaction and task allocation, it does 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 a set of team tasks, task assignments to team members, and the task performance, are fully observable for all team members.

### Explanation procedures

In our experiments, agents provide explanations to users about assignment decisions. There are two different general assignment types during the experiment, the first being an exploration of the strengths and weaknesses of the human teammates in the categories of the tasks, and the next being an assignment made based on the performance of the human teammate vs. the performance of the agent. The latter is presented as either contrastive positive or contrastive negative semantics. At the beginning of the assignment, during the exploration phase, human teammates are presented

with the following message from the agent as an explanation: "I chose this task assignment to explore your strengths and weaknesses". They are then given a set of all task types to explore the human users strengths and weaknesses to rate the initial confidence the agent uses for task assignments. The following presents our experiments for the two different groups, positive and negative contrastive, and their explanation in terms of **A**,**B**,**C**, and **D** where they are replaced with the assignment tasks mentioned in Section 3.

**Assignment Semantics - Constrastive Positive:** In this experiment, contrastive positive explanations, meaning explanations that focus more on what is given to the human teammate rather than what is taken away as an assignment, are given to users by the agent in the following semantic: "I am assigning you **A** Tasks and **B** Tasks because you are better at them compared to **C** Tasks and **D** Tasks". This is positive in that it focuses on what is given based on the users' better skill set in the task categories.

**Assignment Semantics - Contrastive Negative:** In this experiment, Contrastive Negative Explanations, meaning explanations that focus more on what is taken away as an assignment to the Human teammate, are given to users by the agent in the following semantic: "I am not assigning you **C** Tasks and **D** Tasks because you are worse at them compared to **A** Tasks and **B** Tasks." This is negative in that it focuses on the lower skills of the users and the tasks not being assigned.

## Evaluation Metrics

User Performance: A task allocated to a team member is completed successfully or a failure is reported. User performance is measured as the percentage of successful completion of assigned tasks over all episodes: $\sum_{e=1}^{E} \sum_{y_j \in M} \mu_{i,y_j,e}$.

Understandability: Human participants are asked to rate how much they understand the rationale of the agent's actions on a 5-point Likert scale. We use three survey items to measure it. The following is a sample survey question: *"The rationale for task assignments were understandable."* These questions were based on the questions presented in (Shin 2021).

Trust/Confidence: Human participants were asked to rate their confidence and trust in agent allocation decisions on a 5-point Likert scale. We use three survey items to measure it. The following is a sample survey question *"The agent's task allocation decisions makes me confident in its abilities."*. These questions were based on the survey questions presented in (Ehsan et al. 2019a).

Satisfaction: Human participants were asked to rate their satisfaction with the agent teammate on a 5-point Likert scale. We use three survey items to measure it. The following is a sample survey question *"I was satisfied with the experience of using my agent teammate to complete tasks."*.

Explanatory Power: Human participants with explanations, both negative and positive, were asked to rate the explanatory power of the explanations(Ehsan et al. 2019a) on a 5-point Likert scale. This is a rating of how adequate the explanations were in their purpose of providing understandability. We use three survey items to measure it. The following is a sample survey question *"The agent's communications adequately explained the its decisions."*. These survey questions were inspired from the surveys presented in both (Ehsan et al. 2019a) and (Mualla et al. 2022).

Motivation: Human participants with explanations, both negative and positive, were asked to rate how the explanation affected their motivation throughout the episode on a 5-point Likert scale. We used three survey items to measure it. The following is a sample survey question *"Explanations motivated me to work more carefully."*.

# Collaborative Human-Agent Taskboard (CHATboard)

To evaluate the above hypotheses, we needed a domain that encapsulates the following characteristics:

• Team tasks with significant variation in the level of expertise in the general population. Larger variability would allow more space for team adaptation and human satisfaction with teamwork. We should also have the latitude to easily and believably configure the 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 operate effectively 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. 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 collaboration within human teams, to facilitate a human team member perceiving an agent as a distinct team member. Blackboards have also been used effectively in agent teams as a common repository for information exchange between agents (Hayes-Roth 1985). 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 of team information, allowing team members to share and view information through these boards.

## Experimental configurations

We created 32 ($r = 8$) task item instances for each episode, and total number of interactions is four, $E = 4$. The con-

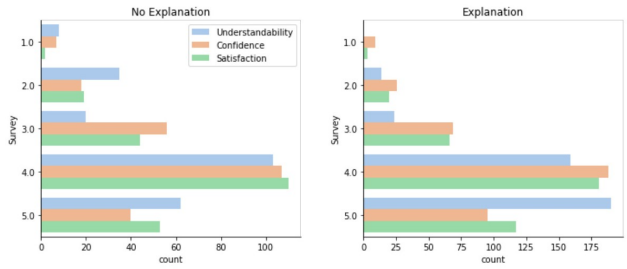


Figure 4: Histogram of human ratings of agent-specific factors with explanations (Right) and no explanations (Left).

fidence levels are stated in a range [1,100], which is then scaled by the agent internally to 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 the capabilities of the team.

The final number of participants recruited for each condition is as follows: 76 for no explanations, 64 for positive conditions, and 65 for negative conditions, as recommended for a medium-sized effect (Brinkman 2009). We use a between-subject experimental design, and each team is randomly assigned 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. We incorporate random comprehension attention checks to ensure the fidelity of the result (Hauser, Paolacci, and Chandler 2019). Participants receive a bonus payment based on performance.

## Experimental Results

We compare results for no-explanation, contrastive positive and negative explanation conditions with respect to their effects on different factors: motivation, explanatory power, understandability, confidence, and satisfaction with agent.

MANOVA test analysis shows that there is a statistically significant difference between no explanation and explanations conditions on factors (understandability, confidence, satisfaction), F = 10.9, p $< 0.01$. We also find a statistically significant effect between no explanation, positive and negative explanations on factors (understandability,confidence,satisfaction), F = 6.8, p $< 0.01$.

**Understandability:** We compare no explanation and explanation conditions with respect to understandability. We found that understandability is higher when there is explanations (M = 4.36, SD = 0.6) than not (M = 3.77, SD = 0.9), and the difference is statistically significant, $t = 4.6, p < 0.01$. Histograms of different human ratings for Understandability and other Agent-Specific Factors are presented in Figure 4. From the two plots, we see that for the explanation case, has a higher frequency of high values for the explanation case in comparison to the no explanation case. **H1a** *is supported.*

We compare the understandability of no-explanation, positive and negative explanations. When comparing positive and negative with no explanation, positive ones (M = 4.30, SD = 0.6) are higher than no explanations (M = 3.77, SD = 0.9), and the difference is statistically significant, $t = 3.7, p < 0.01$, and negative ones (M = 4.42, SD = 0.6) are little higher than no explanations (M = 3.77, SD = 0.9), and the difference is statistically significant, $t = 4.6, p > 0.05$. We also found that the understandability is slightly higher with negative explanations (M = 4.42, SD = 0.6) than positive ones (M = 4.30, SD = 0.6), and the difference is not statistically significant, $t = 0.98, p > 0.05$. **H1b** *is supported.*

Table 1: Factors results for each explanation condition.

| Factors \ XAI | No Exp | | Neg Exp | | Pos Exp | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Motivation | - | - | 3.99 | 0.9 | **4.03** | 0.7 |
| Explanatory Power | - | - | 4.14 | 0.8 | **4.15** | 0.7 |
| Understandability | 3.77 | 0.9 | **4.42** | 0.6 | 4.30 | 0.6 |
| Confidence | 3.68 | 0.8 | 3.79 | 0.9 | **3.94** | 0.8 |
| Satisfaction | 3.85 | 0.8 | 3.87 | 0.8 | **4.14** | 0.6 |
| User Performance | 77.3 | 11.8 | **80.6** | 10.8 | 76.9 | 5.0 |

**Confidence:** We then analyze the effect of the explanation on the confidence of the participants in the agent. We found that confidence is higher when there is explanations (M = 3.86, SD = 0.8) than not (M = 3.68, SD = 0.8), and the difference is not statistically significant, $t = 1.4, p > 0.05$.

When comparing confidence with positive and negative explanations with no explanation, positive ones (M = 3.94, SD = 0.8) are higher than no explanations (M = 3.68, SD = 0.8), although the difference is not statistically significant, $t = 1.8, p > 0.05$, and negative explanations (M = 3.79, SD = 0.9) are little higher than no explanations (M = 3.68, SD = 0.8), although the difference is not statistically significant, $t = 0.7, p > 0.05$. **H2a** *is not supported.*

We also found that confidence is slightly higher with positive explanations (M = 3.94, SD = 0.8) than negative ones (M = 3.79, SD = 0.9), though the difference is not statistically significant, $t = 0.95, p > 0.05$. **H2b** *is not supported.*

**Satisfaction:** We analyze the effect of the explanation on the satisfaction of the participants with the agent. We find satisfaction slightly higher when there is explanations (M = 4.01, SD = 0.7) than not (M = 3.85, SD = 0.8), although the difference is not statistically significant, $t = 1.3, p > 0.05$.

When comparing satisfaction in positive and negative with no explanation, positive ones (M = 4.14, SD = 0.6) are higher than no explanations (M = 3.85, SD = 0.8), and the difference is statistically significant, $t = 2.3, p < 0.05$, while negative ones (M = 3.87, SD = 0.8) are little higher than no explanations (M = 3.85, SD = 0.8) but the difference is not statistically different, $t = 0.1, p > 0.05$. Due to only partial statistical significance, **H3a** *is partially supported.*

We also found that satisfaction is higher with positive explanations (M = 4.14, SD = 0.6) than negative ones (M = 3.87, SD = 0.6), and the difference is borderline statistically significant, $t = 1.9, p = 0.05$. **H3b** *is sup-*

*ported.*

**Motivation:** When comparing motivation with positive and negative explanations, positive ones (M =4.03, SD = 0.7) are similar to negative ones (M =3.99, SD = 0.9), and the difference is not statistically significant, $t = 0.2, p > 0.05$. **H4** *is not supported.*

**Explanatory power:** Similarly, when comparing explanatory power in positive and negative explanations, positive ones (M =4.15, SD = 0.8) are similar to negative ones (M =4.14, SD = 0.8), and the difference is not statistically significant, $t = 0.09, p > 0.05$. **H5** *is supported.*

**Performance:** We then analyze the team when no-explanations, positive and negative explanations are provided. We found team performance to be little higher when explanations (M = 78.7, SD = 8.6) are provided than not (M = 77.3, SD = 10.8) and the difference is not statistically significant, $t = 0.8, p > 0.05$. **H6a** *is not supported.*

When comparing user performance in positive and negative with no-explanation, negative ones (M = 80.6, SD = 10.6) are higher than no explanations (M = 77.3, SD = 10.8), and the difference is not statistically significant, $t = 1.6, p > 0.05$, and positive ones (M = 76.9, SD = 5.0) are slightly lower than no explanations (M = 77.3, SD = 10.8), and the difference is not statistically different, $t = 0.3, p > 0.05$.

We also find that the performance is higher for negative explanations(M = 80.6, SD = 10.6) than positive ones (M = 76.9, SD = 5.0), and the difference is statistically significant, $t = 2.4, p < 0.05$. **H6b** *is not supported.*

## Discussion

Human-agent teams are increasingly used in our lives, and we need to understand how to facilitate collaboration within these teams. We focus on a relatively unexplored area where human and agent teammates collaborate to allocate tasks. Previous work has found that agent allocators outperform their human counterparts, but human satisfaction with such an allocation process was low. In this paper, we strive to understand how explanations of the task allocation process in ad hoc human-agent teams affect human teammate's task performance and satisfaction. Agent allocators can allocate tasks with or without providing an explanation of its decision. Humans expect certain characteristics for explanations (Miller 2019), for example, they are contrastive. In our work task-allocating agents provide either positive and negative contrastive explanations or no explanation at all.

Although humans improved their ratings for agents when the latter provides explanations for allocation decisions, they are indifferent to explanation-associated factors, specifically motivation and explanatory power. We conclude that the explanation objectives of explanatory power, understandability, and motivation, is unaffected by the type of explanation. However, human teammate satisfaction, which is based on human perception, is influenced by the explanation type.

We believe that confidence and trust are approximately the same across explanation types because of the wide acceptability of technology in today's world. As the task types are familiar and not of critical importance to human teammates, humans can mostly trust the agent in this domain. Humans can also conjecture the reasoning behind allocations given the performance metrics, and hence trust the agent even without an explanation. We believe that humans would demand explanations in more complex domains with higher incentives (such as healthcare choices) and where the explanatory power would depend on the explanation given. Furthermore, our domain has three task types, which are memory-based or knowledge-based. Even when allocation decision explanations can improve motivation, they may have limited effect on human ability to successfully complete those tasks.

With the observed effects of the explanation types, particularly on understandability and satisfaction, we conclude that explanations add value to human-agent collaboration. Satisfaction with agents improves when explanations are provided. The type of explanation also has an effect: Humans tend to be more satisfied with positive explanations. Higher levels of human-like communication afford factors such as confidence and satisfaction (Ehsan et al. 2019b). Four characteristics of good explanations have been identified (Miller 2019), including being of social nature. Since more sociable explanations are defined as more friendly or likable, we conjecture that positive explanations are rated better because they are more sociable and human-like than negative explanations. These factors were not measured in this study, but are likely to be relevant and we plan to incorporate them into our future work.

## Conclusions and Future Work

The insights gained in this work suggest that explanations can be useful in ad hoc human-agent teams where task allocator role is assigned to agent teammates; In particular, explanations can increase human teammates' rating of their agent teammates. However, human teammates tend to be indifferent to the tone of the explanations. This may be due to the fact that we do not fully develop our explanations based on the criteria presented in Miller's survey on how humans provide explanations (Miller 2019). In future work, we can incorporate the additional criteria listed in that seminal work, such as social or selected explanations.

We plan to incorporate more of the insights from the social sciences, such as embedding social behavior in the agent teammate. The observed increase of human teammate ratings of agent teammates suggests that endowing agent assigned to allocator roles with the ability to provide explanations is a promising direction to pursue and can serve the goal of increasing both satisfaction and performance in ad hoc human-agent teams.

We can also run experiments with human allocator protocols and ask human team members to provide explanations for how they allocated tasks. We can then use learning approaches to construct explanation templates from the data set of human-generated explanations. These templates can be instantiated by agent allocators to provide explanations, which can potentially be more informative and have high explanatory power.

## References


Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* 58:82–115.

Brinkman, W.-P. 2009. Design of a questionnaire instrument. In *Handbook of mobile technology research methods*. Nova Publishers. 31–57.

Ehsan, U.; Tambwekar, P.; Chan, L.; Harrison, B.; and Riedl, M. O. 2019a. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. *CoRR* abs/1901.03729.

Ehsan, U.; Tambwekar, P.; Chan, L.; Harrison, B.; and Riedl, M. O. 2019b. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 263–274.

Genter, K.; Agmon, N.; and 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)*.

Gervits, F.; Thurston, D.; Thielstrom, R.; Fong, T.; Pham, Q.; and Scheutz, M. 2020. Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In *AAMAS*, 429–437.

Hauser, D.; Paolacci, G.; and Chandler, J. 2019. Common concerns with mturk as a participant pool: Evidence and solutions.

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

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Kim, J.; Muise, C.; Shah, A.; Agarwal, S.; and Shah, J. Bayesian inference of linear temporal logic specifications for contrastive explanations.

Mathieu, J. E.; Hollenbeck, J. R.; van Knippenberg, D.; and Ilgen, D. R. 2017. A century of work teams in the journal of applied psychology. *Journal of applied psychology* 102(3):452.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

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

Mualla, Y.; Tchappi, I.; Kampik, T.; Najjar, A.; Calvaresi, D.; Abbas-Turki, A.; Galland, S.; and Nicolle, C. 2022. The quest of parsimonious xai: A human-agent architecture for explanation formulation. *Artificial Intelligence* 302:103573.

Neerincx, M. A.; van der Waa, J.; Kaptein, F.; and van Diggelen, J. 2018. Using perceptual and cognitive explanations for enhanced human-agent team performance. In Harris, D., ed., *Engineering Psychology and Cognitive Ergonomics*, 204–214. Cham: Springer International Publishing.

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

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.; and Roberts, S. J. 2015. Hac-er: A disaster response system based on human-agent collectives. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 533–541. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Robinette, P.; Wagner, A. R.; and Howard, A. M. 2013. Building and maintaining trust between humans and guidance robots in an emergency. In *AAAI Spring Symposium: Trust and Autonomous Systems*, 78–83.

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

Shin, D. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies* 146:102551.