'Why didn't you allocate this task to them?' Negotiation-Aware Task Allocation and Contrastive Explanation Generation

Zahra Zahedi,[*] **Sailik Sengupta,**[*] **Subbarao Kambhampati**

CIDSE, Arizona State University, USA

zzahedi, sailiks, rao@asu.edu

Abstract

Task-allocation is an important problem in multi-agent systems. It becomes more challenging when the team-members are humans with imperfect knowledge about their teammates' costs and the overall performance metric. While distributed task-allocation methods lets the team-members engage in iterative dialog to reach a consensus, the process can take a considerable amount of time and communication. On the other hand, a centralized method that simply outputs an allocation may result in discontented human team-members who, due to their imperfect knowledge and limited computation capabilities, perceive the allocation to be unfair. To address these challenges, we propose a centralized Artificial Intelligence Task Allocation (AITA) that simulates a negotiation and produces a negotiation-aware task allocation that is fair. If a team-member is unhappy with the proposed allocation, we allow them to question the proposed allocation using a counterfactual. By using parts of the simulated negotiation, we are able to provide contrastive explanations that providing minimum information about other's cost to refute their foil. With human studies, we show that (1) the allocation proposed using our method does indeed appear fair to the majority, and (2) when a counterfactual is raised, explanations generated are easy to comprehend and convincing. Finally, we empirically study the effect of different kinds of incompleteness on the explanation-length and find that underestimation of a teammate's costs often increases it.

Introduction

Whether it be assigning teachers to classes (Kraus et al. 2019), or employees (nurses) to tasks (wards/shifts) (Warner and Prawda 1972), task allocation is essential for the smooth function of human-human teams. In the context of indivisible tasks, the goal of task allocation is to assign individual agents of a team to a subset of tasks such that a pre-defined set of metrics are optimized. For example, in the context of assigning teachers/professors to classes, an allocation may, besides the sole consideration of skill-sets, need to respect each instructor's capacity and timing constraints.

When the cost-information about all the team members and a performance measure is known upfront, one can capture the trade-off between some notion of social welfare

(such as fairness, envy-free, etc.) and team efficiency (or common rewards) (Bertsimas, Farias, and Trichakis 2012). In a distributed setting, agents may have to negotiate back-and-forth to arrive at a final allocation (Saha and Sen 2007). In the negotiation, agents can either choose to accept an allocation proposed by other agents or reject it; upon rejection, agents can propose an alternative allocation that is more profitable for them and, given their knowledge, acceptable to others. While distributed negotiation-based allocations will at least keep the agents happy with their lot (since they got what they negotiated for), it tends to have two major drawbacks, especially when the agents are humans. First, an agent may not be fully aware of their teammates' costs and the performance metrics, resulting in the need for iteratively sharing cost information. Second, the process can be time-consuming and can increase the human's cognitive overload, leading to sub-optimal solutions.

In contrast, a centralized allocation can be efficient, but will certainly be contested by disgruntled agents who, given their incomplete knowledge, may not consider it to be fair. Thus, a centralized system needs to be ready to provide the user with an explanation. As discussed in (Kraus et al. 2019), such explanations can aid in increasing people's satisfaction (Bradley and Sparks 2009) and maintain the system's acceptability (Miller 2018). In a multi-agent environment such as task allocation, providing explanations is considered to be both important and a difficult problem (Kraus et al. 2019).

To address these challenges, we blend aspects of both the (centralized and distributed) approaches and propose an Artificial Intelligence-powered Task Allocator (AITA). Our system (1) uses a centralized allocation algorithm patterned after negotiation to come up with an allocation that explicitly accounts for the costs of the individual agents and overall performance, and (2) can provide contrastive explanation when a proposed allocation is contested using a counterfactual. We assume AITA is aware of all the individual costs and the overall performance costs[1], and use of a negotiation-based mechanism for coming up with a fair or negotiation-aware allocation helps reuse the inference process to pro-

---

[*]indicates equal contribution.

[1]The methods proposed work even when this assumption is relaxed and AITA's proposed allocation and explanation initiate a conversation to elicit human preferences that were not specified upfront. We plan to consider this longitudinal aspect in the future.
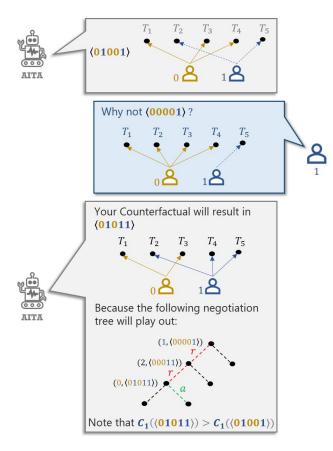
Figure 1: AI Task Allocator (AITA) comes up with a negotiation-aware allocation for a set of human agents. A dissatisfied human agent question AITA with a counterfactual allocation. AITA then explains why its proposed allocation is better than the counterfactual allocation.

vide contrastive explanations. Our explanation has two desirable properties. First, the negotiation-tree based explanation by AITA has a graphical form that effectively distills relevant pieces from a large amount of information (see Figure 1); this is seen as a convenient way to explain information in multi-agent environments (Kraus et al. 2019). Second, the explanation, given it is closely tied to the inference process, acts as a certificate that guarantees fairness[2] to the human (i.e. no other allocation could have been more profitable for them while being acceptable to others). While works like (Kraus et al. 2019) recognize the need for such negotiation-aware and contestable allocation systems, their paper is mostly aspirational. To the best of our knowledge, we are the first to provide concrete algorithms for the generation and explanation of allocation decisions.

Earlier works have taken a purely engineering approach to generating explanations for AI-based decision making. Unfortunately, the claimed benefits do not hold water when gauged via human subject studies (Papenmeier, Englebi-

---
[2]Fairness in this paper implies that the individual views it as fair to them rather than social fairness

enne, and Seifert 2019). To address this, we conduct human studies in three different task-allocation scenarios and show that the allocations proposed by AITA are deemed fair by the majority of subjects. Users who questioned the fairness of AITA's allocations, upon being explained, found it understandable and convincing in two control cases. Further, we consider an approximate version of the negotiation-based algorithm for larger task allocation domains and, via numerical simulation, show how underestimation of a teammate's costs and different aspects of incompleteness affect explanation length.

Related Works

In this section, we situate our work in the landscape of multi-agent task allocation and model reconciliation explanations.

Task allocation   As mentioned above, challenges in task allocation are either solved using centralized or distributed approaches; the later are more considerate of incomplete information settings. Centralized approaches often model the allocation problem as a combinatorial auction and show that maximizing the total sum of bids produce good allocations (Hunsberger and Grosz 2000; Cramton 2006). In human teams, members often have incomplete knowledge about fellow team members and overall performance costs. Thus, a proposed allocation may not seem reasonable to the human. Given existing centralized approaches, there is no way for the human to initiate dialog. Moreover, regardless of optimality or accuracy problem, a centralized decision making system should be contestable in order to engender trust in users. On the other hand, distributed methods allow agents to autonomously negotiate and agree on local deals (Chevaleyre, Endriss, and Maudet 2010; Chevaleyre et al. 2006), finally to reach a consensus– an allocation that is Pareto Optimal (Brams and Taylor 1996). Along these lines, an array of work exists that devise distributed online methods to find such Pareto optimal solutions (Saha and Sen 2007; Endriss et al. 2006, 2003). Beyond negotiation mechanisms, works have also explored the use of bargaining games to model bilateral (i.e. two-agent) negotiation scenarios (Erlich, Hazon, and Kraus 2018; Fatima, Kraus, and Wooldridge 2014; Peled, Kraus et al. 2013). However, we argue that the use of these negotiation methods, often used in distributed settings, by a centralized agent can help is coming up with (1) a fair allocation and (2) an inference method, the later can then be leveraged to provide contrastive explanations to humans who are unhappy with the proposed allocation.

Model Reconciliation Explanations   Human-aware AI systems are often expected to be able to explain their decisions. A good overview about the different kinds of explanations, and what makes them more comprehensible and effective, can be found in (Miller 2018). To generate the various types of explanations, several computation methods exist in AI, ranging from explaining decisions of machine learning systems (Ribeiro, Singh, and Guestrin 2016; Melis and Jaakkola 2018) to explaining sequential decisions by planners (Chakraborti, Sreedharan, and Kambhampati 2020). The need for explanations in our work arises out of

the incomplete knowledge the human has about their team-mates and the team's performance metric. We subscribe the idea of model reconciliation as explanations and, similar to (Sreedharan, Srivastava, and Kambhampati 2018), enable AITA to come up with contrastive explanation when the human deems the proposed allocation as unfair and provides a counterfactual. The counterfactual generation process allows one to make meaningful restrictions about the human's computational capabilities, in contrast to previous works that humans are expected to come up with optimal solutions (Endriss et al. 2003, 2006).

Problem Formulation

Task allocation problems are categorized as mixed-motive situations for humans (especially in setting that agents are ought to fulfill their tasks and cannot dismiss it). They are cooperative in forming a team but in getting the assignment they are selfish and considering their own interest. So, for task assignment humans are selfish but at the same time in order to hold a bargain in the team and keep the team they need to consider other teammates and be cooperative. In other words, they are selfish but there is a bound to their selfishness, and the bound comes so the team will not be broken due to selfishness.

Our task allocation problem can be defined using a 3-tuple $\langle A, T, C \rangle$ where $A = \{0, 1, \ldots, n\}$ and n denotes the AI Task Allocator (AITA) and $0, \ldots n-1$ denotes the n human agents, $T = \{T_1, \ldots, T_m\}$ denotes m indivisible tasks that need to be allocated to the n human agents, and $C = \{C_0, C_1, \ldots C_n\}$ denotes the cost functions of each agent[3]. C_n represents the overall cost metric associated with a task-allocation outcome o (defined below).

For many real world settings, a human may not be fully aware of the utility functions of the other human agents (Saha and Sen 2007). In our setting, a human agent i is only aware of its costs C_i and has noisy information about all the other utility functions $C_j \ \forall j \neq i$ and the performance costs (when $j = n$). We represent i's noisy estimate of j's cost as C_{ij}. For a task t, we denote the human i's cost for that task as $C_i(t)$ their perception of j's cost as $C_{ij}(t)$.

Let O denote the set of allocations and an allocation $o(\in O)$ represent a one-to-many function from the set of humans to tasks; thus, $|O| = n^m$. An outcome o can be written as $\langle o_1, o_2, \ldots o_m \rangle$ where each $o_i \in \{0, \ldots, n-1\}$ denotes the human agent performing task i. Further, let us denote the set of tasks allocated to a human i, given allocation o, as $T_i = \{j : o_j = i\}$. For any allocation $o \in O$, there are two types of costs for AITA:

(1) Cost for each human agent i to adhere to o. In our setting, we consider this cost as $C_i(o) = \Sigma_{j \in T_i} C_i(j)$.

(2) An overall performance cost $C_n(o)$.

---

[3]Preferences over tasks or capability of doing tasks are characterized in cost function (eg. cost of infinity for being incapable in doing a task)

Given the incomplete information setting, a human's perception of costs for an allocation o relates to their (true) cost $C_i(o)$, noisy idea of other agent's costs $C_{ij}(o)$, and noisy idea of the overall team's cost $C_{in}(o)$.

Example. Consider a scenario with two humans $\{0, 1\}$ and five tasks $\{t_1, t_2, t_3, t_4, t_5\}$. An allocation outcome can thus be represented as a binary (in general, base-n) string of length five (in general, length m). For example, $\langle 01001 \rangle$ represents a task allocation in which agent 0 performs the three tasks $T_0 = \{t_1, t_3, t_4\}$ and 1 performs the remaining two tasks $T_1 = \{t_2, t_5\}$. The true cost for human 0 is $C_0(\langle 01001 \rangle) = C_0(t_1) + C_0(t_3) + C_0(t_4)$, while the true cost for 1 is $C_1(t_2) + C_1(t_5)$.

Negotiation Tree A negotiation between agents can be represented as a tree whose nodes represent a two-tuple (i, o) where $i \in A$ is the agent who proposes outcome o as the final-allocation. In each node of the tree, all other agents $j \in A \setminus i$ can choose to either *accept* or *reject* the allocation offer o. If any of them choose to reject o, in a turn-based fashion[4], the next agent $i + 1$ makes an offer o' that is (1) not an offer previously seen in the tree (represented by the set $O_{parents}$, and (2) is optimal in regards to agent $i + 1$'s cost among the remaining offers $O \setminus O_{parents}$. This creates the new child $(i+1, o')$ and the tree progresses either until all agents *accept* the offer or all outcomes are exhausted. Note that in the worst case, the negotiation tree can consist of n^m nodes, each corresponding to one allocation in O. Each negotiation step, represented as a child in the tree, increases the time needed to reach a final task-allocation. Hence, similar to (Baliga and Serrano 1995), we consider a discount factor (given we talk about costs as opposed to utilities) as we progress down the negotiation tree. Although we defined what happens when an offer is rejected, we did not define the criteria for rejection. The condition for rejection or acceptance of an allocation o can be defined as follows.

$$\begin{cases} accept \ o & \text{if} \quad C_i(o) \leq C_i(O^i_{fair}) \\ reject \ o & \text{o.w.} \end{cases}$$

where O^i_{fair} represents a *fair allocation* as per agent i given its knowledge about $C_i, C_{ij}(\forall i \neq j)$, and C_{in} (the latter two being inaccurate). We now define a fair allocation means in our setting, followed by how one can find it.

Proposing a Negotiation-Aware Fair Allocation

In this section, we first formally define a fair allocation followed by how one can computationally find one. We conclude the section with an interpretation of *fair allocation* in terms of the agent's cost functions.

Fair Allocation: An allocation is considered fair by all agents iff, upon negotiation, all the agents are willing to accept it. Formally, an acceptable allocation at step s of the

---

[4]Note that there is an ordering on agents offering allocation (upon rejection by any of the agents). AITA offers first, followed by each team member in some order and then it continues in a round-robin fashion.

negotiation, denoted as $O_{fair}(s)$, has the following properties:

1. All agents believe that allocations at a later step of the negotiation will result in a higher cost for them.

$$\forall i, \ \forall s' > s \quad C_i(O(s')) > C_i(O_{fair}(s))$$

2. All allocations offered by agent i at step s'' before t, denoted as $O_{opt}^i(s'')$, is rejected at least by one other agent. The *opt* in the subscript indicates that the allocation $O_{opt}^i(s'')$ at step s'' has the optimal cost for agent i at step s''. Formally,

$$\forall s'' < s, \ \exists j \neq i, \quad C_j(O_{opt}^i(s'')) > C_j(O_{fair}(s))$$

We now describe how AITA finds a fair allocation.

Fair Allocation Search The negotiation process to find a fair allocation can be viewed as an sequential bargaining game. At each period of the bargaining game, an agent offers an allocation in a round-robin fashion. If this allocation is accepted by all agents, each agent incur a cost corresponding to the tasks they have to accomplish in the allocation proposed (while AITA incurs the team's performance cost). Upon rejection (by even a single agent), the game moves to the next period. Finding the optimal offer (or allocation in such settings) needs to first enumerate all the periods of the sequential bargaining game. In our case, this implies constructing an allocation enumeration tree, i.e. similar to the negotiation tree but considers what happens if all proposed allocations were rejected. In the generation of this allocation enumeration tree, we assume the human agents, owing to limited computational capabilities, can only reason about a subset of the remaining allocations. While any subset selection function can be used in all the algorithms presented in this paper, we will describe a particular one in the next section.

Given that the sequential bargaining game represents an extensive form game, the concept of Nash Equilibrium allows for non-credible threats. In such settings a more refined concept of Subgame Perfect Equilibrium is desired (Osborne et al. 2004). We first define a sub-game and then, the notion of a Sub-game Perfect Equilibrium.

Sub-game: *After any non-terminal history, the part of the game that remains to be played (in our context, the allocations not yet proposed) constitute the sub-game.*

Subgame Perfect Equilibrium (SPE): *A strategy profile s^* is the SPE of a perfect-information extensive form game if for every agent i and every history h (after which i has to take an action), the agent i cannot reduce its cost by choosing a different action, say a_i, not in s^* while other agents stick to their respective actions. If $o_h(s^*)$ denotes the outcome of history h when players take actions dictated by s^*, then $C_i(o_h(s^*)) \leq C_i(o_h(a_i, s_{-i}^*))$.*

Given the allocation enumeration tree, we can use the notion of *backward induction* to find the optimal move for all agents in every sub-game (Osborne et al. 2004). We first start from the leaf of the tree with a sub-tree of length one. We then keep moving towards the root, keeping in mind the best strategy of each agent (and the allocation it leads to). We

claim that if we repeat this procedure until we reach the root, we will find a fair allocation. To guarantee this, we prove two results– (1) an SPE always exists and can be found by our procedure and (2) the SPE returned is a fair allocation.

Lemma *There exists a non-empty set of SPE. An element of this set is returned by the backward induction procedure.*

Proof Sketch. Note that the backward induction procedure always returns a strategy profile; in the worst case, it corresponds to the last allocation offered in the allocation enumeration tree. Each agent selects the optimal action at each node of the allocation enumeration tree. As each node represents the history of actions taken till that point, any allocation node returned by backward induction (resultant of optimal actions taken by agents), represents a strategy profile that is an SPE by definition. Thus, an SPE always exists. □

Corollary *The allocation returned by backward induction procedure is a fair allocation.*

Proof Sketch. A proof by contradiction shows that if the allocation returned is not a fair allocation, then it is not the SPE of the negotiation game, contradicting *Lemma*. □

Interpreting Negotiation-aware Allocations in Terms of Individual Costs Let us denote the optimal allocation for an agent i as O_{opt}^i. Note that for any multi-agent setting, $C_i(O_{opt}^i) = 0$ because in the optimal case, all tasks are allocated to the other agents $j(\neq i)$. Clearly, the negotiation process prunes out these solutions because at least one other agent $j(\neq i)$ will reject this solution.[5] Now let us denote the allocation obtained by the backward induction procedure as O_{SPE}. For an agent i, this allocation is $\Delta = C_i(O_{SPE}) - C_i(O_{opt}^i) = C_i(O_{SPE})$ away from the optimal solution they could have obtained (Δ is showing the bound on human's selfishness in mixed-motive situations). Given that all agents either (1) make an offer based on that is the least cost for them at a particular step of negotiation or (2) reject offers based on their belief of getting a lower cost solution in the future, the fair allocation O_{SPE} is guaranteed to be the closest to their optimal solution accepted by all other agents.

Explaining a Proposed Allocation

In this section, we first introduce the notion of a human's counterfactual that arises due to incomplete information and limited inferential capabilities. We then present the notion of a contrastive explanation in this context and show that given the worst-case counterfactual allocation, there always exists such an explanation.

Counterfactual Allocation AITA is truthful and will not exploit ignorance (human's limit knowledge about other teammates) in favor of anything. So, to find the negotiation-aware allocation, it assumes and acts as everybody knows everything and finds a fair allocation, in respect to this assump-

---

[5]Note that this assumption might not hold if the # indivisible tasks < # agents as there exists an agent i with no tasks.
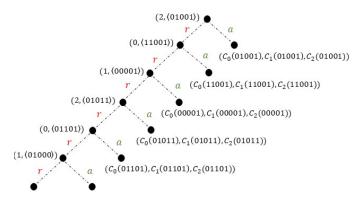
Figure 2: Negotiation tree enumerated by agent 1 to come up with the best counterfactual when offered $o = \langle 01001 \rangle$.

tion. Thus, given this assumption nobody should question the allocation given by AITA. But because the assumption doesn't hold in practice, the human may question the allocation. So, in scenarios where the human has (1) access to all the true costs and (2) computation capabilities to reason about a full negotiation tree, they will simply realize that AITA's allocation is fair and not need an explanation. Thus, AITA, with the correct information about the costs, came up with a negotiation-aware fair allocation and proposed it to the human. Humans, with noisy estimates[6] of the other's costs and limited computational abilities, may be able to come up with an allocation given o that they think is lower cost for them and will be *accepted* by all other players. Note that human may assume the centralized agent may make inadvertent errors but not deliberate misallocation and there is no consideration about deception, misuse or irrationality in this work.

Formally, we define the notion of an optimal counterfactual as follows.

The optimal counterfactual *for a human agent i is an alternative allocation o', given AITA's proposed allocation o, that has the following properties.*

1. *o' is in the set of allocations regarding their limited computational capability*   *(this implies that $o \neq o'$)*

2. *$C_i(o) > C_i(o')$*   *(i has lower cost in o')*

3. *o' is an SPE in the allocation enumeration tree made from allocations from o given their computational capability.*

We now state assumptions made about the computational capabilities of a human.

Limited Computational Capabilities: Subset Selection Function We assume that given a particular allocation outcome $o = \langle o_1, \ldots o_j \rangle$, a human will only consider outcomes o' where only one task in o is allocated to a different human j. In the context of our example, given allocation $\langle 010 \rangle$, the human can only consider the three allocations $\langle 011 \rangle$, $\langle 000 \rangle$

_____

[6]noisy estimates of other agents are not needed to be known by AITA in order to find a fair allocation or generate explanation.

and $\langle 110 \rangle$; outcomes are one Hamming distance away. With this assumption in place, a human is considered capable of reasoning about a negotiation tree with $m * (n - 1)$ allocations (as opposed to n^m) in the worst case[7].

Figure 2 shows the negotiation tree an agent used to come up with the optimal counterfactual allocation. An SPE of this tree provides the optimal counterfactual the human can offer (note that when SPE$= o$, no counterfactual exists).

Explanation as a Negotiation Tree

We show that when an optimal counterfactual o_i exist for a human agent $i \in \{0, \ldots, n-1\}$, AITA can come up with an effective explanation that refutes it, i.e. explains why o is a better solution for i than their counterfactual o_i.

We thus propose to describe a negotiation tree, which starts with the counterfactual o_i at its root and excludes the original allocation o, as an explanation. This differs from the human's negotiation tree because it uses the actual costs as opposed to using the noisy cost estimates. Further, AITA, with no limits on computation capabilities, can propose allocations that are not bound by the the subset selection function. We finally show that an SPE in this tree results in an SPE that cannot yield a lower cost for the human i than o. At that point, we expect a rational human to be convinced that o is a better allocation than the counterfactual o_i. Note that a large allocation problem does not imply a long explanation (i.e. a negotiation tree of longer path from root to lowest leaf). In turn, a larger problem does not necessarily make an explanation verbose. We can now define what an explanation is in our case.

Explanation. *An explanation is a negotiation tree with true costs that shows the counterfactual allocation o_i will result in a final allocation $\hat{o}_i$ such that $C_i(\hat{o}_i) \geq C_i(o)$.*

Even before describing how this looks in the context of our example, we need to ensure one important aspect– given a counterfactual o' against o, an explanation always exists.

Proposition *Given allocation o (the fair allocation offered by AITA) and a counterfactual allocation o_i (offered by i), there will always exist an explanation.*

Proof Sketch: We showcase a proof by contradiction; consider an explanation does not exist. It would imply that there exists $\hat{o}_i$ that reduces human i's cost (i.e. $C_i(o) \geq C_i(\hat{o}_i)$) and is accepted by all other players after negotiation. By construction, o was a fair allocation and thus, if a human was able to reduce its costs without increasing another agent's cost, all agents will not have accepted $\hat{o}_i$. As o is also the resultant of a sub-game perfect equilibrium strategy of the allocation enumeration tree with true costs AITA would have chosen $\hat{o}_i$. Given AITA chose o, there cannot exist such a $\hat{o}_i$. □

_____

[7]As specified above, other ways to limit the computational capability of a human can be factored into the backward induction algorithm. Due to the lack of literature on how compute abilities may be relaxed in a task-allocation setting, we consider 1-edit distance as a starting point.
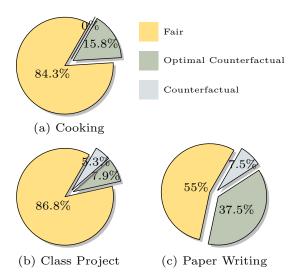
Figure 3: Options selected by participants for the three different domains used in the human studies.

Experimental Results

In this section, we try to examine the fairness of the proposed allocations and the effectiveness of the contrastive explanations generated by AITA. For this, we consider two different human study experiments. In one setting, if a user selects an allocation is unfair, we ask them to rate our explanation and two other baselines (relative case). In the other case, we simply provide them our explanation (absolute case). We then consider the effect of different kinds of inaccuracies (about costs) on the explanation length for a well-known project management domain (Certa et al. 2009).

Human Subject Studies We briefly describe the study setup for the two human studies.

Relative Case In this setting, we consider two different task allocation scenarios. The first one considers cooking an important meal at a restaurant with three tasks and two teammates– the chef and the sous-chef. The second one considers dividing five tasks associated with a class project between a senior grad and a junior grad/undergrad student. In the first setting, the human subject plays the role of a sous-chef and are told they are less efficient at cooking meals and may produce a meal of low overall quality (the performance cost is interpreted as the customer ratings seen so far). In the second setting, the human subject fills in the role of the senior student who is more efficient at literature review, writing reports and can yield a better quality project (as per the grading scheme of an instructor). We recruited a total of 38 participants of whom 54.3% were undergraduate and 45.7% were graduate students in Computer Science & Engineering and Industrial Engineering at our university[8]. All of them

---

[8]the fact that college students could actually understand the scenarios and explanations for domains they are not experts in (eg. cooking) or had stakes in (they were not really asked to do the

were made to answer a few filter questions correctly to ensure they fully understood the scenarios.

In the study, we presented the participants with AITA's proposed allocation and counterfactual allocations that adhere to the one-hamming distance subset selection function defined above. This let us consider two and three counterfactual allocations for the cooking and the class project domains respectively.[9] When the human selects a counterfactual, implying that AITA's proposed allocation is unfair, we present them with three explanations. Besides our negotiation-tree based explanation, we showcase two baseline explanations– (1) A *vacuous* explanation that simply states that the human's counterfactual won't be accepted by others and doesn't ensure a good overall performance metric, (2) A *verbose* explanation that provides the cost of all their teammates and the performance metric for all allocations.

Absolute Case Setup In this case, we considered a task allocation scenario where two research students– a senior researcher and a junior researcher– are writing a paper and the three tasks relate to working on different parts of the paper. In this setting, we gathered data from 40 graduate students. Similar to the previous case, the subjects have to select whether the AITA's proposed allocation if fair or select between either of the two counterfactual allocations (each adhering to the subset selection constraint). In contrast to the previous case, upon selecting a counterfactual, the subject is only presented with the negotiation-tree explanation.

Results. Across the three different domains, as shown in Figure 3, we noticed that a majority of the participants selected that AITA's allocation is fair. This shows that the formally defined metric of fairness based on negotiation (negotiation-aware allocation) does indeed appear fair to humans. Given that a set of the participants felt that the allocation is unfair and demand counterfactual allocations, we (1) question the validity of other fairness metrics defined technically but never tested via human subject studies and (2) establish the need for explanations in task-allocation settings. We then noticed that the next highest selected option was the optimal counterfactual that was calculated using the SPE mechanism over the human's negotiation tree (that is generated assuming the human's computational capabilities are limited). For the cooking and paper writing domains, the result was statistically significant, highlighting that our computational methods to come up with the optimal counterfactual is in line with how humans would come up with a counterfactual in these task allocation domains. The least selected option was the other sub-optimal counterfactual allocations.

For participants who selected the explanations in the relative case, they were asked to rate the comprehensibility and convincing power of the three explanations provided to them on a scale of $1-5$. The results of the experiments are shown in Table 1. In this setting, the vacuous explanation, a state-

---

tasks) proves the generalizability of our results obtained from the human-subjects evaluations

[9]A detailed description of the domains and the study can be found in the supplementary material

Table 1: Human study scores for our negotiation-tree based explanations compared to the two baselines.

Domain	Explanation	Understandable $(1-5)$	Convincing $(1-5)$
Cooking	Vacuous	4.5	2.33
	Verbose	4.3	4
	Neg-tree	4.5	4
Class Project	Vacuous	4.4	2.8
	Verbose	4.2	3.4
	Neg-tree	3.8	4.4

Table 2: Agent's true costs for completing a task.

	t_1	t_2	t_3	t_4	t_5
1	0.3	0.5	0.4	0.077	0.8
2	0.4	0.7	0.077	0.49	0.13

ment that the human's allocation won't be accepted by others and does not guarantee better performance compared to AITA's proposed allocation, was easy to understand but was least convincing in both the cooking domain and the class project domain. The verbose explanations, that provided the human with all the possible costs it wasn't aware of, was the least understandable in the cooking domain and more understandable than the negotiation-tree based explanation in the class project domain. It was regarded more convincing in the cooking domain, which had fewer tasks and hence fewer number of costs provided to the human, and appeared less convincing for the class project domain as the number of tasks (and in turn the number of allocations) increased. The negotiation tree based explanation appeared to be the most understandable explanation in the cooking domain and the most convincing one in both domains. In the absolute control case, where the two baselines were removed, instead of using a rating from $1-5$, we used a Likert scale as ensuring (1) the scale of different human participants might be different and (2) without baselines, they might not adhere to a consistent scoring scheme. In this setting, the negotiation tree was judged to be *understandable* and *moderately convincing* on average. This is similar to the understandable scores seen in the class project domain but a little less convincing. It shows that when the explanation is presented independently without baselines, the humans believe their might be more convincing alternatives. Hence, the negotiation-tree based explanations for multi-agent task allocation marks a good first step towards an exciting direction for future research (Kraus et al. 2019).

Impact of Noise on Explanations For this study, we use the project management scenario from (Certa et al. 2009) in which human resources are allocated to *R&D* projects. We modify the original setting in three ways. First, we consider two and four human agents instead of eight for assigning the five projects, allowing a total of $2^5 = 32$ possible allocations. It allows for explanations of reasonable length where

allocation can be represented as 5-bit binary strings (see Figure 1). Second, we only consider the skill aspect, ignoring the learning abilities of individuals and social aspects of an allocation. This was mostly done because we could not confidently specify a relative prioritization of the three. We use the skill to measure the time needed, and in turn the cost, for completing a project (more the time needed, more the cost). There are a total of $2 * 5 = 10$ actual costs, 5 for each human (shown in Table 2), and 10 additional costs representing the noisy perception of one human's cost by their teamamte. Third, we consider an aggregate metric that considers the time taken by the two humans to complete all the tasks. Corresponding to each allocation, there are 32 (true) costs for team performance shown below (not enumerated here for space considerations). With these cost, as shown in Figure 1, the negotiation-aware allocation is $\langle 01001 \rangle$, the optimal counterfactual for agent 1 is $\langle 00001 \rangle$ which is revoked by AITA using an explanation tree of length three.

Impact of Norm-bounded Noise. The actual cost C_i of each human i as a vector of length m. A noisy assumption can be represented by another vector situated ϵ $(l2)$ distance away. By controlling ϵ, we can adjust the magnitude of noise a human has. In Figure 4, we plot the effect of noise on the average explanation length. The noisy cost vectors are sampled from the l_2 norm ball withing ϵ radius scaled by highest cost in the actual cost vectors (Calafiore, Dabbene, and Tempo 1998).[10] The y-axis indicates the length of the replay negotiation tree shown to the human. Even though the maximum length of explanation could be $31 (2^5 - 1)$, we saw the maximum explanation length was 8. Given that every noise injected results in a different scenario, we average the explanation length across ten runs (indicated by the solid lines). We also plot the additive variance (indicated by the dotted lines). The high variance on the negative side (not plotted) is a result of the cases where either (or both) of the human(s) human team members didn't have an optimal counterfactual and thus, the explanation length was zero.

We initially hypothesized, based on intuition, that an increase in the amount of noise will result in a longer explanation. The curve in red (with ∘) is indicative that this is not true. To understand why this happens, we classified noise into two types– Optimistic Noise (ON) and Pessimistic Noise (PN)– representing the scenarios when a human agent overestimates or under-estimates the cost of the other human agents for performing a task. When a human overestimates the cost of others, it realizes edits to a proposed allocation will lead to a higher cost for other agents who will thus reject it. Thus, optimal counterfactual ceases to exist and thus, explanations have length zero (reducing the average length). In the case of PN, the human underestimates the cost of teammates. Thus, upon being given an allocation, they often feel this is unfair and find an optimal counterfactual demanding explanations. As random noise is a combination of both ON and PN (overestimates costs of some humans for particular tasks but underestimates their cost for other tasks etc.),

---

[10]We clip the norm-ball to be in the positive quadrant as negative costs are meaningless in our setting.
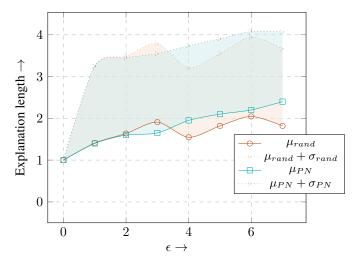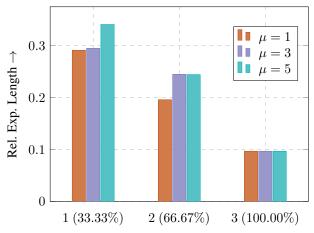
Figure 4: Mean length of explanations as the amount of noise added to the actual costs increases.

Figure 5: As the number of agents about whom a human has complete knowledge (co-workers whose costs you know) increases, the mean length of explanations decreases.

the increase in the length of explanations is counteracted by zero length explanations. Hence, in expectation, we do not see an increase in explanation length as we increase the random noise magnitude. As per this understanding, when we increase ϵ and only allow for PN, we clearly see an increase in the mean explanation length (shown in green).

When $\epsilon = 0$, there is no noise added to the costs, i.e. the humans have complete knowledge about the team's and the other agent's costs. Yet, due to limited computational capabilities, a human may still come up with a counterfactual that demands explanation. Hence, we observe a mean explanation length of 1 even for zero noise. This should not be surprising because, when coming up with a foil, a human only reasons in the space of $n * (m - 1)$ allocations (instead of n^m allocations).

Incompleteness about a sub-set of agents. In many real-world scenarios, an agent may have complete knowledge about some of their team-mates but noisy assumptions about others. To study the impact of such incompleteness, we considered a modified scenario project-management domain (Certa et al. 2009) with four tasks and four humans. We then choose to vary the size of the sub-set about whom a human has complete knowledge. In Fig. 5, we plot the mean length of explanations, depending upon the subset size about whom the human has complete knowledge. On the x-axis, we plot the size of the subset and on the y-axis, we show the relative explanation length that equals to explanation length divided by the longest explanation ($4^4 = 256$) we can have in this setting. We consider five runs for each sub-set size and only pessimistic noise (that ensures a high probability of having a counterfactual and thus, needing explanations). We notice as the number of individuals about whom a human has complete knowledge increases, the mean relative explanation length (times the max explanation length) decreases uniformly across the different magnitude of noise μ. Even when a human has complete knowledge about all other

agent's costs, happens whenever the size of the sub-set is $n - 1$ (three in this case), it may still have some incompleteness about the team's performance costs. Added with limited computational capabilities (to search in the space of 16 allocations), they might still be able to come up with counterfactual; in turn, needing explanations. Thus, the third set of bar graphs (corresponding to the label 3 (100.00%) on the x-axis) has a mean of ≈ 0.1 relative explanation length.

Conclusion

In this paper, we considered a task-allocation problem where a centralized AI Task Allocator (AITA) comes up with a negotiation-aware fair allocation using a simulated negotiation based approach, which are popular in distributed task allocation settings, for a team of humans. When the humans have limited computational capability and incomplete information about all costs except their own, they may be dissatisfied with AITA's proposed allocation and question AITA using counterfactual allocations that they believe are fairer. We show that in such cases, AITA is able to come up with a negotiation tree that (1) representative of the inference methodology used and (2) explains that if the counterfactual was considered, it would result in final allocation that is worse-off than the one proposed. With human studies, we show that the negotiation-aware allocation appears as fair to majority of humans while for the others, our explanations are *understandable* and *convincing*. We also perform experiments to show that when agents either overestimate the cost of other agents or have accurate information about more agents, the average length of explanations decreases.

References

Baliga, S.; and Serrano, R. 1995. Multilateral bargaining with imperfect information. *Journal of Economic Theory* 67(2): 578–589.

Bertsimas, D.; Farias, V. F.; and Trichakis, N. 2012. On the efficiency-fairness trade-off. *Management Science* 58(12): 2234–2250.

Bradley, G. L.; and Sparks, B. A. 2009. Dealing with service failures: The use of explanations. *Journal of Travel & Tourism Marketing* 26(2): 129–143.

Brams, S. J.; and Taylor, A. D. 1996. *Fair Division: From cake-cutting to dispute resolution*. Cambridge University Press.

Calafiore, G.; Dabbene, F.; and Tempo, R. 1998. Uniform sample generation in l/sub p/balls for probabilistic robustness analysis. In *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No. 98CH36171)*, volume 3, 3335–3340. IEEE.

Certa, A.; Enea, M.; Galante, G.; and Manuela La Fata, C. 2009. Multi-objective human resources allocation in R&D projects planning. *International Journal of Production Research* 47(13): 3503–3523.

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2020. The Emerging Landscape of Explainable AI Planning and Decision Making. *arXiv preprint arXiv:2002.11697* .

Chevaleyre, Y.; Dunne, P. E.; Endriss, U.; Lang, J.; Lemaitre, M.; Maudet, N.; Padget, J.; Phelps, S.; Rodriguez-Aguilar, J. A.; and Sousa, P. 2006. Issues in multiagent resource allocation. *Informatica* 30(1).

Chevaleyre, Y.; Endriss, U.; and Maudet, N. 2010. Simple negotiation schemes for agents with simple preferences: Sufficiency, necessity and maximality. *Autonomous Agents and Multi-Agent Systems* 20(2): 234–259.

Cramton, P. 2006. Introduction to combinatorial auctions. P. Cramton, Y. Shoham, R. Steinberg, eds., Combinatorial Auctions.

Endriss, U.; Maudet, N.; Sadri, F.; and Toni, F. 2003. On optimal outcomes of negotiations over resources. In *AAMAS*, volume 3, 177–184.

Endriss, U.; Maudet, N.; Sadri, F.; and Toni, F. 2006. Negotiating socially optimal allocations of resources. *Journal of artificial intelligence research* .

Erlich, S.; Hazon, N.; and Kraus, S. 2018. Negotiation strategies for agents with ordinal preferences. *arXiv preprint arXiv:1805.00913* .

Fatima, S.; Kraus, S.; and Wooldridge, M. 2014. The negotiation game. *IEEE Intelligent Systems* 29(5): 57–61.

Hunsberger, L.; and Grosz, B. J. 2000. A combinatorial auction for collaborative planning. In *Proceedings fourth international conference on multiagent systems*, 151–158. IEEE.

Kraus, S.; Azaria, A.; Fiosina, J.; Greve, M.; Hazon, N.; Kolbe, L.; Lembcke, T.-B.; Müller, J. P.; Schleibaum, S.; and Vollrath, M. 2019. AI for Explaining Decisions in Multi-Agent Environments. *arXiv preprint arXiv:1910.04404* .

Melis, D. A.; and Jaakkola, T. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, 7775–7784.

Miller, T. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* .

Osborne, M. J.; et al. 2004. *An introduction to game theory*, volume 3. Oxford university press New York.

Papenmeier, A.; Englebienne, G.; and Seifert, C. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* .

Peled, N.; Kraus, S.; et al. 2013. An agent design for repeated negotiation and information revelation with people. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

Saha, S.; and Sen, S. 2007. An Efficient Protocol for Negotiation over Multiple Indivisible Resources. In *IJCAI*, volume 7, 1494–1499.

Sreedharan, S.; Srivastava, S.; and Kambhampati, S. 2018. Hierarchical Expertise Level Modeling for User Specific Contrastive Explanations. In *IJCAI*, 4829–4836.

Warner, D. M.; and Prawda, J. 1972. A mathematical programming model for scheduling nursing personnel in a hospital. *Management Science* 19(4-part-1): 411–422.