'Why not give this work to them?' Explaining AI-Moderated Task-Allocation Outcomes using Negotiation Trees

Zahra Zahedi[*†], **Sailik Sengupta**[*], **Subbarao Kambhampati**
CIDSE, Arizona State University, USA
{zzahedi, sailiks, rao}@asu.edu

Abstract

The problem of multi-agent task allocation arises in a variety of scenarios involving human teams. In many such settings, human teammates may act with selfish motives and try to minimize their cost metrics. In the absence of (1) complete knowledge about the reward of other agents and (2) the team's overall cost associated with a particular allocation outcome, distributed algorithms can only arrive at sub-optimal solutions within a reasonable amount of time. To address these challenges, we introduce the notion of an AI Task Allocator (AITA) that, with complete knowledge, comes up with fair allocations that strike a balance between the individual human costs and the team's performance cost. To ensure that AITA is explicable to the humans, we allow each human agent to question AITA's proposed allocation with counterfactual allocations. In response, we design AITA to provide a replay negotiation tree that acts as an explanation showing why the counterfactual allocation, with the correct costs, will eventually result in a sub-optimal allocation. This explanation also updates a human's incomplete knowledge about their teammate's and the team's actual costs. We then investigate whether humans are (1) able to understand the explanations provided and (2) convinced by it using human factor studies. Finally, we show the effect of various kinds of incompleteness on the length of explanations. We conclude that underestimation of other's costs often leads to the need for explanations and in turn, longer explanations on average.

1 Introduction

Allocation problems are ubiquitous in various settings ranging from industrial processes to financial markets. In these problems, a set of agents try to agree on allocation of tasks in a way that is good for them and respects some notion of social welfare such as fairness, envy-free, etc. Methods used to reach an allocation outcome comprise of either centralized or distributed algorithms where agents share (complete or partial) information about their rewards pertaining to a set of tasks. In these processes, an agent may either choose to accept or reject solutions proposed by another agent, in turn proposing solutions they think are profitable to them and acceptable to the others. Such interactions are termed as *negotiation*.

In this work, we relax the assumption that human agents have the cognitive capabilities to reason through an entire negotiation process, taking into account incomplete/noisy information about other agents' costs and the team's performance metric. In such settings, an Artificially Intelligent Task-Allocation (AITA) agent that is aware of (1) the costs for all individual agents and (2) the team performance metric can aid in coming up with a fair allocation (defined formally later). In doing so, AITA champions the cause for improving team performance metrics, a variable that neither of the (selfish) humans may care about. This setup resembles scenarios in industrial warehouses where a centralized system assigns tasks to personnel, who are expected to follow them.

Given that AITA takes away the need for the human agents to perform a negotiation process, a human agent may not be necessarily pleased with the proposed allocation because, with their incomplete knowledge and limited computational capability, they might believe that the proposed allocation is not profitable for them. Thus, we allow the human agents to provide a counterfactual allocation that they think would reduce their cost while being accepted by the other agents. AITA can then replay a part of the negotiation tree, along with the true costs of the other agents, to justify why the proposed allocation was fair. We show that our explanation always acts as a certificate to guarantee fairness for a human (i.e. no other allocation could have been more profitable for them while being acceptable to others).

In essence, AITA can provide optimal task allocation and, when questioned using counterfactual (or foils), can explain why the current allocation is the better one by using (1) reward information unknown to the explaniee and (2) allocation that would result due to the negotiation that would occur if the counterfactual was considered. As shown in Fig. 1, the entire process can be visualized as a conversation between AITA and an agent.

[*]indicates equal contribution.
[†]Contact author.

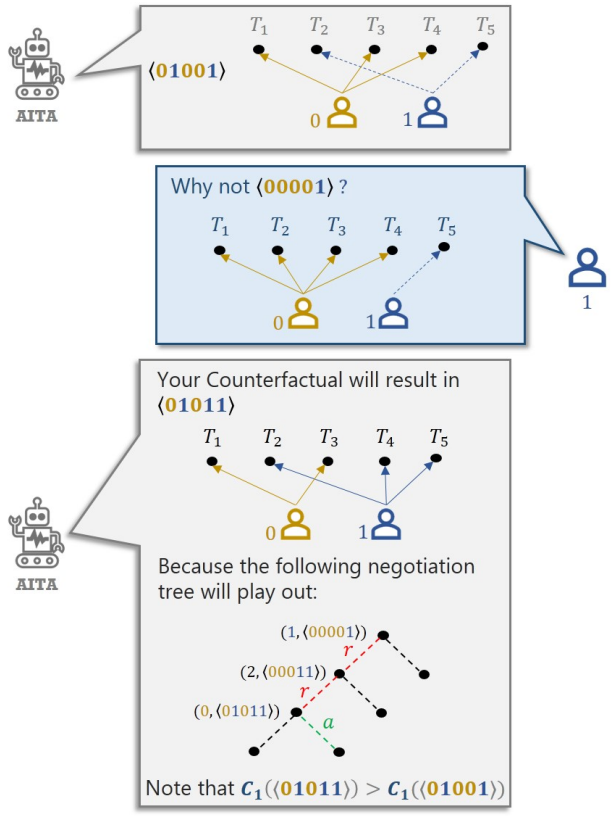

Figure 1: AI Task Allocation (AITA) agent comes up with a task-allocation for a set of human agents. A human agent can come up with a counterfactual allocation. AITA can then explain why its proposed allocation is better than the counterfactual allocation.

2 Related Works

Our work draws inspiration from a set of works in multi-agent task allocation and, as already states, explanations in human-aware AI. Prior works in multi-agent task allocation can be divided into centralized and distributed approaches; the latter more relevant in incomplete information settings. The former methods focus on combinatorial auctions that have a centralized process for coming up with a final resource allocation which maximizes the sum of the prices associated with the bids [Hunsberger and Grosz, 2000; Cramton, 2006]. In the distributed allocation settings, the agents autonomously negotiate on resources to agree on local deals [Chevaleyre *et al.*, 2010; 2006]. These works seek to arrive at an allocation (or often called, a deal) that guarantees Pareto optimality [Brams and Taylor, 1996]. There exists an array of work on finding such Pareto optimal allocations in such settings [Saha and Sen, 2007; Endriss *et al.*, 2006; 2003]. Similar to our work, there exists a spectrum of work that consider the use of bargaining games for modeling bilateral (i.e. two-agent) negotiation scenarios [Erlich *et al.*, 2018; Fatima *et al.*, 2014; Peled *et al.*, 2013]. In contrast to all these works, we focus on (1) having a centralized entity that comes up with fair allocations and (2) how such an entity can provide explanations to human agents who, in the presence of incomplete knowledge and limited computational capabilities, may come up with a counterfactual allocations that they believe is more optimal for them than the proposed fair allocation.

There also exists an array of work on coming up with explanations in human-aware AI systems– ranging from a general overview of what explanations should look like and when they are effective [Miller, 2018] to explanations that explain decisions of machine learning systems [Melis and Jaakkola, 2018; Ribeiro *et al.*, 2016] or plans generated by an automated planners [Chakraborti *et al.*, 2017; Sreedharan *et al.*, 2017; 2018; Borgo *et al.*, 2018]. Our work is more closely related to the latter work that views explanations as model reconciliation, where the model of the AI-agent is assumed to be error-free while the explainee's model might be imprecise or incomplete. In such cases, given a plan, the human might not be convinced that it is an optimal one. Thus, they can ask for explanations (at times, based on an alternative foil [Sreedharan *et al.*, 2018]). In our setting, the human agents have an incorrect model of the other agent's cost and the team performance metric. Thus, given an allocation, they might be able to think of small edits (counter-factual allocation) they believe will make the proposed allocation better for them while not making it worse for other agents. AITA will replay the negotiation that would take place if the counterfactual is considered, showing to the human that with the true costs (only for the allocations in the replay tree) the counterfactual will result in a more sub-optimal allocation for them. As opposed to existing work that expects a human is able to come up with optimal plans in their head [Endriss *et al.*, 2003; 2006], we put reasonable restrictions on the human's computational capabilities.

3 Problem Formulation

Our task allocation problem can be defined using a 3-tuple $\langle A, T, C\rangle$ where $A = \{0, 1, \ldots, n\}$ where 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 tasks (or indivisible resources) 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. 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 ($\forall\ i > 0$) is only aware of its costs C_i and has noisy information about all the other utility functions $C_j\ \forall\ j \neq i$. We represent i's noisy estimate of j's cost as C_{ij}. Note that because $j \in \{0, \ldots, n\}$, the human also has a noisy estimate of the team's cost. Given a task t, we denote the human i's cost for that task as $C_i(t)$. Similarly, human i's perception of j's cost for performing task t is denoted as $C_{ij}(t)$.

An outcome O is a task allocation such that every task $\in T$ is allocated to exactly one human $\in A \setminus \{0\}$. Given that an outcome is a one-to-many function from the set of human agents to the set of tasks, $|O| = n^m$. Also, an outcome o can be written as $\langle o_1, o_2, \ldots o_m\rangle$ where each $o_i \in \{0, \ldots, n-1\}$ and denotes the human agent performing task i. 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 that can be denoted as a function of the human's cost for performing the individual tasks allocated to them. In our setting, we consider this cost, say for human i, as $C_i(o) = \Sigma_{j \in T_i} C_i(j)$.
(2) a performance cost for the given outcome defined over the allocation outcome O (eg. $C_n(o)$).

Given incomplete information about the rewards of other agents, 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 $C_{in}(o)$.

As an example. consider a scenario with two humans $\{0, 1\}$ and five tasks $\{t_1, t_2, t_3, t_4, t_5\}$ (will be discussed later). 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, the next agent $i + 1$ (if $i + 1 > n, i + 1 = 0$) is asked to make 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 previous works [Baliga and Serrano, 1995], we consider a discount factor (rather a multiplicative 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 do not define the criteria for rejection. As a responder, an agent's strategy is

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

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.

4 Proposing a 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 time t of the negotiation, denoted as $O_{fair}(t)$, has the following properties:
1. All agents believe that allocations at a later time of the negotiation will result in a higher cost for them.

$$\forall i,\ \forall t^{'} > t \quad C_i(O(t^{'})) > C_i(O_{fair}(t))$$

2. All allocations offered by agent i at time t'' before t, denoted as $O^i_{min}(t'')$, is rejected at least by one other agent. The min in the subscript indicates that the allocation $O^i_{min}(t'')$ at time t'' has the least cost for agent i at time t''. Formally,

$$\forall t^{''} < t,\ \exists j \neq i, \quad C_j(O^i_{min}(t^{''})) > C_j(O_{fair}(t))$$

We now describe how an agent can find a fair allocation.

4.1 Finding a Fair Allocation

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.

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 and others, 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 and others, 2004]. In this, 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 results in a fair allocation.

Lemma 1. *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 by the backward induction procedure is not a fair allocation, then it is not the SPE of the negotiation game, contradicting *Lemma 1*.

4.2 Interpreting Fair Allocations

Although we define a fair allocation with respect to the negotiation procedure, we now discuss how it can be interpreted in terms of the cost incurred by an agent. Let us denote the optimal allocation for an agent i as O^i_{opt}. Note that for any multi-agent setting, $C_i(O^i_{opt}) = 0$ because in the optimal case, all tasks are allocated to the other agents $j(\neq i)$. Clearly, the negotiation process prunes out this solution because at least one other agent $j(\neq i)$ will reject this solution.[1]

Let us denote the allocation obtained by the backward induction procedure as O_{SPE}. For an agent i, this allocation is $\Delta = C_i(O_{SPE}) - C_i(O^i_{opt}) = C_i(O_{SPE})$ away from the optimal solution they could have obtained. Given that all agents either (1) make an offer based on that is least cost for them at a particular time 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.

5 Explaining a Proposed Allocation

In this section, we describe the notion of a human's counterfactual allocation. For this, we assume certain inferential capabilities the human has. We describe what a contrastive explanation means in this context and prove that given a counterfactual allocation, there always exists such an explanation.

5.1 Human's Counterfactual Allocation

AITA, with the correct information about the costs, was able to come up with a fair allocation. Now it can offer this as an outcome to the human. Humans who have noisy estimates of the other costs (and limited computational abilities), may come up with an allocation given o, which they think can result in lower cost for them and will be *accepted* by all other players. This notion of a counterfactual can be formally defined as follows.

[1]Note that this assumption might not hold if the number of tasks are less than the number of agents and each task can be done by one agent because in such settings, there will always exist an i who is not allocated any tasks.

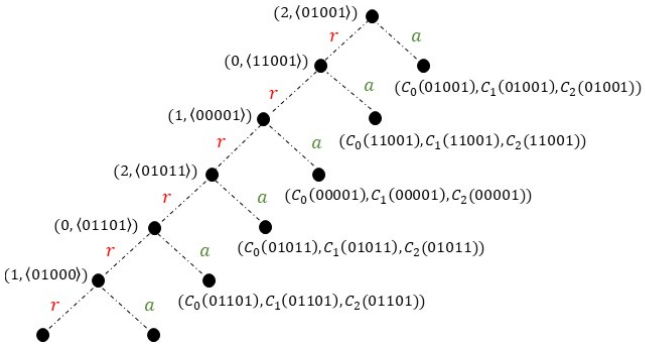

Figure 2: Negotiation tree enumerated by agent 1 to come up with the best counterfactual when offered $o = \langle 01001 \rangle$.

Counterfactual allocation *Given an allocation o, an alternative o' in human i's model is a counterfactual iff*

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 allocation enumeration tree made from allocations from o given their computational capability.*

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

Human Capabilities

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 $\langle 011 \rangle$, $\langle 000 \rangle$ 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[2].

Fig. 2 shows the negotiation tree an agent used to come up with the optimal counterfactual allocation. The reason is similar to saying 'if I instead chose o_1, AITA will reject it because of high team costs and agent 1 in its turn will offer o_2 which I will reject and then AITA will offer ...'. An SPE of this tree provides the best counterfactual (if SPE $\neq o$) the human can offer.

5.2 Explanation as a Negotiation Tree

We now show that whenever there exists an optimal counterfactual o' for at least one of the human agents $i \in \{0, \ldots, n-1\}$, AITA can come up with an effective explanation that shows to the human(s) i that o is a better solution for them than their counterfactual o_i.

We thus, propose to describe a negotiation tree, which starts with the counterfactual at its root and excludes the original allocation o, as an explanation. This differs from the hu-

[2]any way 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.

man's negotiation tree because it uses the actual costs as opposed to using the noisy cost estimates. Also, AITA can provide allocations it chose when asked to offer allocations and these may not necessarily be a one-edit distance away from o_i. 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 allocation they offered. 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.

Lemma 2. *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 $\hat{o}_i$. □

6 Experimental Results

In this section, we consider two kinds of experiments– (1) human factors on a synthetically designed task-allocation domain and (2) the effect various amount of inaccuracies (about costs) have on the explanation length for a well known project management domain [Certa *et al.*, 2009].

6.1 Human Factors

Setup and Study Details The task allocation scenario for the study, keeping in mind that the subjects of our study were only graduate students, was based on allocating three different parts of writing a paper–introduction, literature review and main contribution– to the human (subject) and their co-worker. The subjects were given their cost and only heuristic guidance about their co-worker's cost, who was a senior graduate student. We did not provide them with a noisy estimate to see the assumptions they would organically make about the other cost and the paper quality (the overall team cost). We had a total of 40 graduate students participate in the study.

Results AITA provided them with the (fair) allocation $o = \langle Y, C, Y \rangle$ where Y denotes the task (intro and main contribution) was assigned to the subject and C denotes the task (literature review) allocated to their co-worker. The subjects were first asked if they considered a counterfactual allocation they think would improve their costs while being accepted by others. Among the 3 options given to them, 22 participants felt that the task allocation was fair and said that they did not want to offer a counterfactual, 3 considered $\langle Y, C, C \rangle$ while the remaining 15 considered $\langle C, C, Y \rangle$ as a counterfactual.

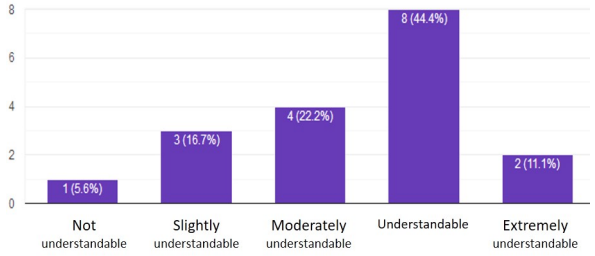

Figure 3: Was the explanation understandable?

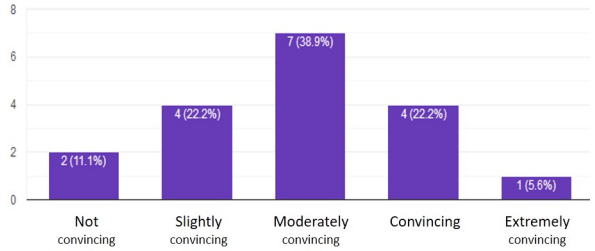

Figure 4: Was the explanation convinving?

People who considered the allocation as fair were asked to explain their answer. We gathered from their subjective response that most participants overestimated the actual cost of their co-worker in performing the literature review and thus, believed that an allocation that imposes any more work on them will be rejected. Interestingly, a couple of participants reasoned that if they were to perform all the three tasks alone, they would be doing (almost) double the work. Thus, with the inherent assumption that their co-worker is equally skilled as them in all the tasks, they felt this was a fair allocation of tasks (as they were allocated $\approx 50\%$ of the total work). Thus, if their estimates about the other agent were not drastically far off than the reality, our idea of a fair allocation seemed to resonate with majority of the participants.

For the 18 participants who felt that there was a counterfactual allocation that improved their costs and would be accepted, we provided them with the replay negotiation tree as an explanation and then asked them two questions– (1) was it understandable? and (2) was it convincing? The responses to there were measured using the Likert scale. *As per the results show in Figure 3 and 4, majority of the participants felt that the explanation was 'understandable' and 'moderately convincing'*. By analysing the response of the six participants who felt the explanation was not or slightly convincing, we saw that most of them failed to understand the cost calculations shown at each node of the negotiation tree. Some felt that the overall paper quality was not being considered (although we showed the overall cost considered by AITA in the explanation). For the others, they suggested that breaking down the cost calculations or augmenting the negotiation tree with a description would convince them more.

6.2 Impact of Noise on Explanations

Scenario Description We use the project management scenario from [Certa *et al.*, 2009] in which human resources (aka

	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

Table 1: Agent's true costs for completing a task.

researchers) have to be allocated to *R&D* projects. We modify the original setting in the following ways. First, we consider the problem with two human agents (instead of eight) while keeping the number of projects (equal to five) to be the same. As there are $2^5 = 32$ possible allocations, it helps us come up with examples of reasonable length. Also, we can represent these allocations by five bit binary strings that are easier to understand. Second, The project allocation problem considers various aspects of a human agent such as skills, learning abilities and social relationships. However, we consider only the skill aspect.[3] We use the skill measure to come up with the time needed for completing a project and make it equal to the agent's cost (more the time needed to complete, more the cost). There are a total of $2 * 5 = 10$ costs, 5 for each human (shown in Table 1). Further, we have 10 additional costs that represent the noisy perception of one human's cost by another human. 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 and limitations of compute capabilities, as shown in Figure 1, the SPE (or the fair 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 about C_i can be represented by another vector situated ϵ distance away using $l2$ distance measure. By controlling ϵ, we can adjust the magnitude of noise each human has.

In Figure 5, we highlight the effect of noise on the explanation length. The x-axis indicates the amount of noise we injected. The noisy cost was sampled from the l_2 norm ball with radius equal to the highest cost in the vector multiplied by ϵ [Calafiore *et al.*, 1998].[4] The y-axis indicates the length of the replay negotiation tree shown to the human, called *explanation length*. Even though the maximum length of explanation could be $31(2^5 - 1)$ we say that explanation length was, at max, eight. 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) to give the reader a sense of the explanation length we show the human. The high variance on the negative side, not plotted, is a result of the cases where either (or both) of the human(s) didn't have a counterfactual; thus, the explanation length was zero.

[3]Although we set the weightage of the learning abilities and social relationships to be zero, a function that can combine the three measures into a real number can also be used for cost calculations.

[4]Given that negative costs don't make sense in our setting, we consider only the non-negative space inside the norm-ball.

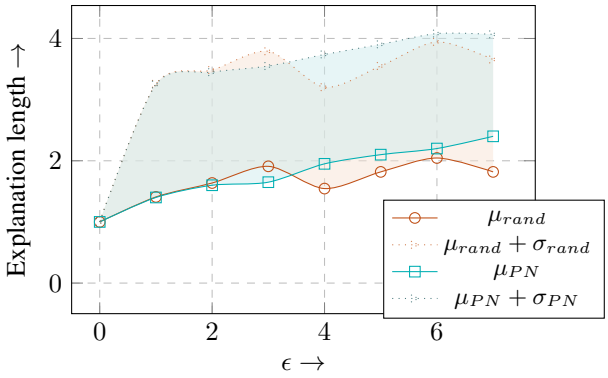

Figure 5: Mean length of explanations as the amount of noise added to the actual costs increases.

We initially hypothesized, appealing to common sense, 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, given an allocation, it realizes any edit will lead to a higher cost for other agents, who will thus reject it. Thus, it becomes harder for them to come up with a good counterfactual; in turn having explanations of length zero (that reduces the average length of explanations). In the case of PN, the human thinks that other humans can easily perform all the tasks, i.e. over-estimates their skill, and in turn underestimates their cost. Thus, upon being given an allocation, they will often find a one-edit allocation that they think is less costly for them and others won't reject; in turn, increasing the average length of 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.), 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 computation capabilities, a human can still come up with a counterfactual that needs explanation. Hence, we see a mean explanation length of 1 even for zero noise.

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. 6, we plot the mean length of explanations, depending upon the subset size about whom

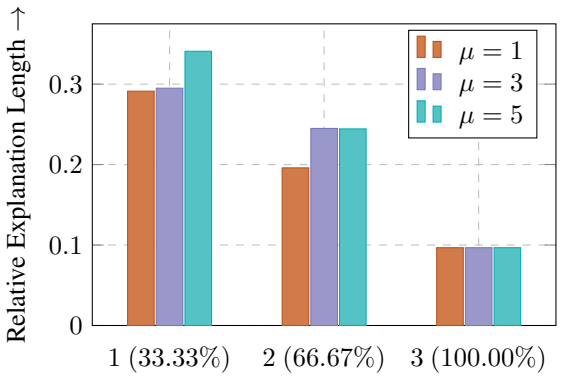

Figure 6: As the number of agents about whom a human has complete knowledge (co-workers whose costs you know) increases, the mean length of explanations decreases.

the human has complete knowledge.

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.

7 Conclusion

In this paper, we considered a task-allocation problem where an AI Task Allocator (AITA) comes up with a fair allocation for a group of humans. When humans have limited computational capability and incomplete information about all other costs than their own, they can come up with counterfactual allocations that they believe are more fair. We show that in such cases, AITA is able to come up with a negotiation tree that explains if the counterfactual is considered, it would result in final allocation that is costlier that the one proposed. With human studies, we show that our notion of fair allocation aligns well with majority of humans while for the others, our explanations are *understandable* and *moderately 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.

Acknowledgments. This research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF-19-2-0006, NSF grants 1936997 (C-ACCEL), 1844325, NASA grant NNX17AD06G, and a JP Morgan AI Faculty Research grant. Sailik Sengupta is supported by the IBM Ph.D. Fellowship.

References

[Baliga and Serrano, 1995] Sandeep Baliga and Roberto Serrano. Multilateral bargaining with imperfect information. *Journal of Economic Theory*, 67(2):578–589, 1995.

[Borgo *et al.*, 2018] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. Towards providing explanations for ai planner decisions. *arXiv preprint arXiv:1810.06338*, 2018.

[Brams and Taylor, 1996] Steven J Brams and Alan D Taylor. *Fair Division: From cake-cutting to dispute resolution*. Cambridge University Press, 1996.

[Calafiore *et al.*, 1998] G Calafiore, F Dabbene, and R Tempo. 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, pages 3335–3340. IEEE, 1998.

[Certa *et al.*, 2009] Antonella Certa, Mario Enea, Giacomo Galante, and Concetta Manuela La Fata. Multi-objective human resources allocation in r&d projects planning. *International Journal of Production Research*, 47(13):3503–3523, 2009.

[Chakraborti *et al.*, 2017] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. ijcai international joint conference on artificial intelligence (2017), 156–163, 2017.

[Chevaleyre *et al.*, 2006] Yann Chevaleyre, Paul E Dunne, Ulle Endriss, Jérôme Lang, Michel Lemaitre, Nicolas Maudet, Julian Padget, Steve Phelps, Juan A Rodriguez-Aguilar, and Paulo Sousa. Issues in multiagent resource allocation. *Informatica*, 30(1), 2006.

[Chevaleyre *et al.*, 2010] Yann Chevaleyre, Ulle Endriss, and Nicolas Maudet. Simple negotiation schemes for agents with simple preferences: Sufficiency, necessity and maximality. *Autonomous Agents and Multi-Agent Systems*, 20(2):234–259, 2010.

[Cramton, 2006] P Cramton. Introduction to combinatorial auctions. p. cramton, y. shoham, r. steinberg, eds., combinatorial auctions, 2006.

[Endriss *et al.*, 2003] Ulrich Endriss, Nicolas Maudet, Fariba Sadri, and Francesca Toni. On optimal outcomes of negotiations over resources. In *AAMAS*, volume 3, pages 177–184, 2003.

[Endriss *et al.*, 2006] Ulrich Endriss, Nicolas Maudet, Fariba Sadri, and Francesca Toni. Negotiating socially optimal allocations of resources. *Journal of artificial intelligence research*, 2006.

[Erlich *et al.*, 2018] Sefi Erlich, Noam Hazon, and Sarit Kraus. Negotiation strategies for agents with ordinal preferences. *arXiv preprint arXiv:1805.00913*, 2018.

[Fatima *et al.*, 2014] Shaheen Fatima, Sarit Kraus, and Michael Wooldridge. The negotiation game. *IEEE Intelligent Systems*, 29(5):57–61, 2014.

[Hunsberger and Grosz, 2000] Luke Hunsberger and Barbara J Grosz. A combinatorial auction for collaborative planning. In *Proceedings fourth international conference on multiagent systems*, pages 151–158. IEEE, 2000.

[Melis and Jaakkola, 2018] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pages 7775–7784, 2018.

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

[Osborne and others, 2004] Martin J Osborne et al. *An introduction to game theory*, volume 3. Oxford university press New York, 2004.

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

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.

[Saha and Sen, 2007] Sabyasachi Saha and Sandip Sen. An efficient protocol for negotiation over multiple indivisible resources. In *IJCAI*, volume 7, pages 1494–1499, 2007.

[Sreedharan *et al.*, 2017] Sarath Sreedharan, Subbarao Kambhampati, et al. Explanations as model reconciliationa multi-agent perspective. In *2017 AAAI Fall Symposium Series*, 2017.

[Sreedharan *et al.*, 2018] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. Hierarchical expertise level modeling for user specific contrastive explanations. In *IJCAI*, pages 4829–4836, 2018.