'Why Didn't You Allocate This Task to Them?' Negotiation-Aware Task Allocation and Contrastive Explanation Generation

Zahra Zahedi[1], **Sailik Sengupta**[2*], **Subbarao Kambhampati**[1]
[1]SCAI, Arizona State University
[2] AWS AI Labs
zzahedi@asu.edu, sailiks@amazon.com, rao@asu.edu

Abstract

In this work, we design an Artificially Intelligent Task Allocator (AITA) that proposes a task allocation for a team of humans. A key property of this allocation is that when an agent with imperfect knowledge (about their teammate's costs and/or the team's performance metric) contests the allocation with a counterfactual, a contrastive explanation can always be provided to showcase why the proposed allocation is better than the proposed counterfactual. For this, we consider a negotiation process that produces a negotiation-aware task allocation and, when contested, leverages a negotiation tree to provide a contrastive explanation. With human subject studies, we show that the proposed allocation indeed appears fair to a majority of participants and, when not, the explanations generated are judged as convincing and easy to comprehend.

Introduction

Whether it be assigning teachers to classes (Kraus et al. 2020), 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. 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, still 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 for human negotiators. 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 deem it to be unacceptable. Thus, a centralized system needs to be ready to provide the user with an explanation. As discussed in (Kraus et al. 2020), 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 challenging problem (Kraus et al. 2020).

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] Use of a negotiation-based mechanism for coming up with a negotiation-aware explicable allocation helps reuse the inference process to provide contrastive explanations. Our explanations have 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. 2020). Second, the explanation, given it is closely tied to the inference process, acts as a certificate that guarantees explicability 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. 2020) recognize the need for such negotiation-aware and contestable allocation systems, their work is mostly aspirational. To the best of our knowl-

*The work was done while at Arizona State University

[1]The proposed methods 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.

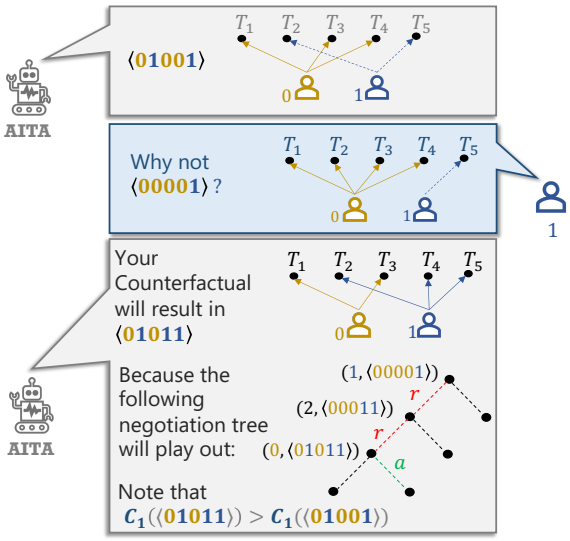

Figure 1: AI Task Allocator (AITA) comes up with a negotiation-aware explicable allocation $\langle 01001 \rangle$ for a set of two humans– 0 and 1. In this allocation, human 0 is assigned tasks 1, 3 and 4 and agent 1 is assigned tasks 2 and 5. A dissatisfied human 1 questions AITA with a counterfactual allocation $\langle 00001 \rangle$, where he/she just needs to do task 5 (they believe task 5 is much more difficult and will take similar effort compared to doing all the 4 others). AITA then explains why the original proposed allocation (i.e. $\langle 01001 \rangle$) is better than the counterfactual allocation (i.e. $\langle 00001 \rangle$). The graph of the negotiation tree can be given as a dialogue "if human 1 proposes the allocation $\langle 00001 \rangle$, it will be rejected and AITA will offer $\langle 00011 \rangle$, which will then be rejected and human 0 will propose a counter offer $\langle 01011 \rangle$ which will then will have to be accepted by all. This final allocation would have a higher cost for you (human 1) than the first proposed allocation. Hence, the counterfactual allocation will eventually result in worse-off allocation for human 1.

edge, we are the first to provide concrete algorithms for the generation and explanation of allocation decisions. To evaluate our work, we conduct human studies in three different task allocation scenarios and show that the allocations proposed by AITA are perceived as fair by the majority of subjects. Users who questioned 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, explicable decision making and model reconciliation explanations.

Task Allocation Task allocation solutions typically involve either centralized or distributed approaches; the later are more considerate of incomplete information settings. In centralized approaches, the allocation is often modeled as a combinatorial auction (Hunsberger and Grosz 2000; Cramton 2006). Note that individual members in teams often have incomplete knowledge about fellow team members and overall performance costs. Thus, a proposed allocation may not seem reasonable to a human. Given existing centralized approaches, there is no way for a disgruntled human to initiate dialog. Essentially, regardless of optimality or accuracy, a centralized decision making system should be contestable. 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; Saha and Sen 2007; Endriss et al. 2006, 2003). Beyond negotiation mechanisms, works have also explored 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).

Explicable Decision Making Effective human-aware AI systems should generate decisions that are aligned with human expectations. In this regard, explicability is defined based on the alignment of an AI agent's decision to a human's expectations. For planning problems, explicability is calculated in several ways, such as syntactic difference, learned labeling function, and cost difference (Kulkarni et al. 2019; Zhang et al. 2017; Olmo, Sengupta, and Kambhampati 2020; Sreedharan et al. 2020). While explicability in team-based planning has been examined (Sengupta, Chakraborti, and Kambhampati 2018), explicability in situations where decisions matter to multiple individuals with unique preferences and expectations of each individual have not been fully considered. Our work addresses this by developing negotiation-aware explicable allocation that is close to every human's optimal allocation and Pareto optimal.

Model Reconciliation Explanations AI systems should be able to explain their decisions (Miller 2018), which can be achieved by various computational methods (Ribeiro, Singh, and Guestrin 2016; Melis and Jaakkola 2018; Rosenfeld and Richardson 2019; Anjomshoae et al. 2019; Mualla et al. 2022; Chakraborti, Sreedharan, and Kambhampati 2020). The literature emphasizes the need for causal explanations and contestability in decision making systems (Lyons, Velloso, and Miller 2021; van der Waa et al. 2021; Kluttz, Kohli, and Mulligan 2022; Aler Tubella et al. 2020; Madumal et al. 2020). While prior studies have explored explanations for scheduling problems through unsatisfied constraints (Agrawal, Yelamanchili, and Chien 2020; Bertolucci et al. 2021; Pozanco et al. 2022) or argumentation (Čyras et al. 2019), our work uniquely provides negotiation-aware explanation through model reconciliation for a multi-human task-allocation problem. The need for explanations stems from humans' incomplete knowledge about their team and performance metrics. We subscribe to the idea of model reconciliation as explanations and allows AITA to provide contrastive explanations when proposed task-allocation is contested with a counterfactual (Sreedharan, Srivastava, and Kambhampati 2018).

Problem Formulation

Task allocation problems are categorized as mixed-motive situations for humans (especially in setting where 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 $A = \{0, 1, \ldots, n\}$ where n denotes the AI Task Allocator (AITA) and $0, \ldots n-1$ denotes the n humans, $T = \{t_1, \ldots, t_m\}$ denotes m indivisible tasks that need to be allocated to the n humans, and $C = \{C_0, C_1, \ldots C_n\}$ denotes the cost functions of each agent (preferences over tasks, skills and capability of doing tasks are characterized in cost function– eg. cost of infinity for being incapable in doing a task). C_n represents the overall performance cost metric associated with a task allocation outcome o (defined below).

For a task t, we denote the human i's cost for that task as $C_i(t)$. 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 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 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 $C_n(o)$.

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[2], 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 with 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). This factor is specifically applied by multiplying it with the costs as we traverse 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.

$$\begin{cases} accept\ o & if\ C_i(o) \leq C_i(O^i_{na\_exp}) \\ reject\ o & otherwise \end{cases} \tag{1}$$

where $O^i_{na\_exp}$ represents a *negotiation-aware explicable allocation* as per agent i.

Proposing a Negotiation-Aware Explicable Allocation

In this section, we first formally define a negotiation-aware explicable allocation followed by how it can be computed.

Definition: Negotiation-Aware Explicable Allocation An allocation is considered explicable by all agents iff, upon negotiation, all the agents are willing to accept it. Formally, an acceptable allocation at step s of the negotiation, denoted as $O_{na\_exp}(s)$, 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.

$$\forall i,\ \forall s^{'} > s \qquad C_i(O(s^{'})) > C_i(O_{na\_exp}(s)) \tag{2}$$

2. All allocations offered by agent i at step s'' before s, denoted as $O^i_{opt}(s'')$, are rejected at least by one other agent. The *opt* in the subscript indicates that the allocation $O^i_{opt}(s'')$ at step s'' has the optimal cost for agent i at step s''. Formally,

$$\forall s^{''} < s,\ \exists j \neq i, \quad C_j(O^i_{opt}(s^{''})) > C_j(O_{na\_exp}(s)) \tag{3}$$

We now describe how AITA finds a negotiation-aware explicable allocation.

Negotiation-Aware Explicable Allocation Search The negotiation process to find an explicable allocation can be viewed as a 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 incurs 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

[2]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.

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 humans, 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 Sub-game 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) constitutes the sub-game.*

Sub-game 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 negotiation-aware explicable 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 an explicable 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 an acceptable allocation.*

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

Given the introduced algorithm, AITA, with the correct information about the costs, and considering the limited computational capability of the humans, came up with a negotiation-aware explicable allocation and proposed it to the human.

Counterfactual Allocation and Explaining a Proposed Allocation

Counterfactual Allocation In scenarios where the human has (1) access to all the true costs of them and others and (2) computation capabilities to reason about a full negotiation tree, they will simply realize that AITA's allocation is explicable and does not need an explanation. However, for many real world settings, a human may not be fully aware of the utility functions of the other humans (Saha and Sen 2007). So in our setting, we formalize this by assuming a human 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 perception of j's cost as $C_{ij}(t)$. 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 performance cost $C_{in}(o)$. Given human i's limited computational abilities, and knowledge about $C_i, C_{ij}(\forall i \neq j)$, and C_{in} (the latter two being inaccurate), $O^i_{na\_exp}$ might not be equal to O_{na_exp} that AITA proposes. Therefore, a counterfactual allocation may raise by human i who may be able to come up with an allocation that they think has 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 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.*

Explanation as a Negotiation Tree We show that when an optimal counterfactual o_i exists for a human $i \in \{0, \ldots, n-1\}$, AITA can come up with an effective explanation that refutes it, i.e. explains how the counterfactual o_i can later result in an acceptable allocation that is not better than o.

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.[3] Further, AITA, with no limits on computation capabilities, can propose allocations that are not bounded by 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 (note that this negotiation tree can be cast as a natural language or dialogue) 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 explicable 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; suppose 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 an explicable allocation and thus, if a human was able to reduce its costs without increasing another agent's cost, all agents will not have accepted o. As $\hat{o}_i$ 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 $\hat{o}_i$. □

An Example of Limited Computational Capabilities: Subset Selection Function An example of subset function can be assuming 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.[4] While there can be other ways to assume subset function that shows human limited computational capability. The described 1-edit distance function will be used in Experimental Results as an assumption for limited computational capability of human.

[3]Noisy estimates of other agents are not needed to be known by AITA in order to generate explanation.

[4]As specified above, other ways to limit the computational capability of a human can be factored into the backward induction algorithm.

Domain	Percentage of selected options		
	Fair	Optimal Counterfactual	Sub-optimal Counterfactual
Cooking	84.2%	15.8%	0%
Class Project	86.4%	7.9%	5.3%
Paper Writing	55%	37.5%	7.5%

Table 1: Options selected by participants for the three different domains used in the human studies.

Experimental Results

In this section, we evaluate our proposed negotiation-aware explicable allocation by assessing its perception among human users. The study asks the human subjects whether AITA's allocation seem fair (as in acceptable) to them. We do admit that the the answers by the human subjects assess their informal/subjective assessment rather than prove any specific formal definition of fairness. We believe that the study does still offer valuable assessment of AITA framework. Additionally, we assess the effectiveness of the contrastive explanations generated by AITA.

In this section, we consider two different human study experiments. In one setting, if a user selects an allocation that 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 . All of them 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.[5]. When the human selects a counterfactual, implying that AITA's proposed allocation is inexplicable, 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 is 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 In Table 1, we see that a majority of the participants selected AITA's allocation as fair across all the three different domains. This shows that our formally defined negotiation-aware explicable allocation does indeed appear fair to majority of participants. Given that a set of the participants felt that the allocation is unfair, we were able to establish a scenario where they desired explanations. Instead of having them calculate a counterfactual, we presented them with options they would might want to use as a counterfactual. We noticed that the highest selected counterfactual was the optimal counterfactual, calculated using the SPE mechanism over the human's negotiation tree in the background (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 are indeed able to come up with the optimal counterfactual that is in line with how humans would come up with a counterfactual. The least selected option was the other sub-optimal counterfactual allocations.

For participants who asked for explanations by providing a counterfactual, we had two settings– relative and absolute. In the relative case, they were asked to rate the comprehensibility and convincing power of the three explanations on a scale of $1-5$. In absolute case, only our explanation was provided to them. We observed that the negotiation tree was judged to be *understandable* and *moderately convincing* on average in the absolute setting. In the relative setting, results in both the cooking and the class project domain show that our explanation is the most convincing one. It is also perceived as the most understandable explanation, but shared the stage with the Vacuous explanation (the average scores and additional metrics are available in the supplementary file).

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

Statistical Significance To further compare our negotiation-tree based explanation with the other two– vacuous and verbose– explanations, we performed one-way ANOVA and a non-parametric Kruskal-Wallis tests with Bonferroni correction. The results of both ANOVA test ($F(2,24) = 8.96$, $p = 0.001$), and Kruskal-Wallis test ($p = 0.003$) show that (1) the three explanations are different in term of convincing power and (2) the negotiation-tree based explanation is the most convincing while the vacuous explanation is the least convincing (rank sum: $172.5 > 140 > 65.5$). However, ANOVA ($F(2,24) = 0.04$, $p = 0.96$) and Kruskal-Wallis ($p = 0.68$) tests show that all the three explanations are not greatly different when it comes to human understanding. So, for understanding, we further performed pair-wise comparison of explanations with pairwise Mann-Whitney tests. Along similar lines, this test also didn't show any significant difference between the pairs ($p = 0.63$, 0.42 and 0.73). Finally, we used the TOST (equivalent test) to evaluate if three explanations are equivalent in terms of understanding. TOST showed that all three explanations are all equal in terms of understanding with 90% confidence interval $(-0.7\ \ 1.2)$ and $(-0.6\ \ 0.8)$ in $(-1.5\ \ 1.5)$. Therefore, we can conclude that while *our negotiation-tree based explanation is the most convincing one*, it is similar to the other in terms of understandable rating by humans.

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 humans 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 allocations 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 (the detail costs in supplementary material), and 10 additional costs representing the noisy perception of one human's cost by their teammate. 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. With these costs, as shown in Figure 1, the negotiation-aware explicable 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 (in l_2 norm) distance away. By controlling ϵ, we can adjust the magnitude of noise a human has. In Figure 2, we plot the effect of noise on the average explanation length. The noisy cost vectors are sampled from the l_2 norm ball within ϵ radius scaled by highest cost in the actual cost vectors (Calafiore, Dabbene,

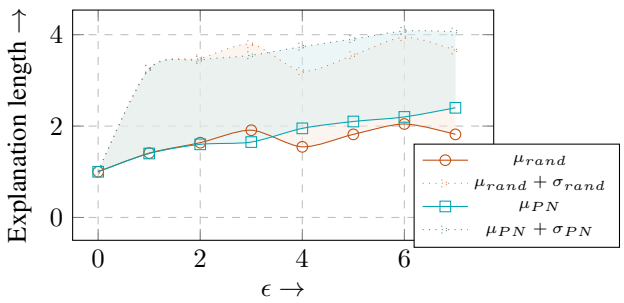

Figure 2: Mean length of explanations as the amount of noise added to the actual costs increases.

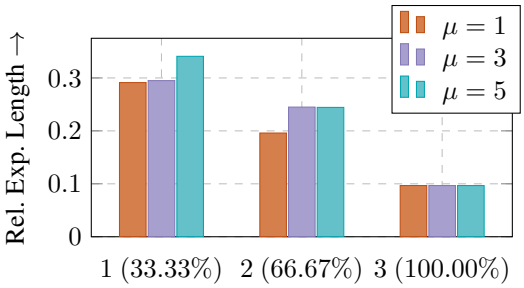

Figure 3: The mean explanation length decreases as the number of co-workers' costs known to a human increases.

and Tempo 1998). 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 overestimates or under-estimates the cost of the other humans 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 inexplicable 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.), 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 subset 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. 3, 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

We explored a task allocation scenario involving a centralized AI Task Allocator (AITA) using simulated negotiation methods for a human team. We examined situations where humans, limited by computational capacity and incomplete cost information, question AITA's allocations using counterfactuals that they believe are explicable. We demonstrated that AITA can generate a negotiation tree to effectively communicate why the counterfactual would yield a less optimal result. Human studies confirmed that the negotiation-aware allocation is perceived as fair by most participants and our explanations are understandable and convincing to the remainder. Our experiments also showed a decrease in explanation lengths when agents either overestimate others' costs or possess more accurate information about them.

Acknowledgments

This research is supported in part by ONR grants N00014-18-1-2442, N14-18-1-2840 and N00014-23-1-2409 and a JP Morgan AI Faculty Research grant to Kambhampati.

References

Agrawal, J.; Yelamanchili, A.; and Chien, S. 2020. Using explainable scheduling for the mars 2020 rover mission. *arXiv preprint arXiv:2011.08733*.

Aler Tubella, A.; Theodorou, A.; Dignum, V.; and Michael, L. 2020. Contestable black boxes. In *International Joint Conference on Rules and Reasoning*, 159–167. Springer.

Anjomshoae, S.; Najjar, A.; Calvaresi, D.; and Främling, K. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

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

Bertolucci, R.; Dodaro, C.; Galatà, G.; Maratea, M.; Porro, I.; and Ricca, F. 2021. Explaining ASP-based Operating Room Schedules. In *IPS-RCRA@ AI* IA*.

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.

Čyras, K.; Letsios, D.; Misener, R.; and Toni, F. 2019. Argumentation for explainable scheduling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2752–2759.

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.

Kluttz, D. N.; Kohli, N.; and Mulligan, D. K. 2022. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In *Ethics of Data and Analytics*, 420–428. Auerbach Publications.

Kraus, S.; Azaria, A.; Fiosina, J.; Greve, M.; Hazon, N.; Kolbe, L.; Lembcke, T.-B.; Muller, J. P.; Schleibaum, S.; and Vollrath, M. 2020. AI for explaining decisions in multiagent environments. 34(09): 13534–13538.

Kulkarni, A.; Zha, Y.; Chakraborti, T.; Vadlamudi, S. G.; Zhang, Y.; and Kambhampati, S. 2019. Explicable planning as minimizing distance from expected behavior. In *AAMAS*.

Lyons, H.; Velloso, E.; and Miller, T. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–25.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2493–2500.

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*.

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.

Olmo, A.; Sengupta, S.; and Kambhampati, S. 2020. Not all failure modes are created equal: Training deep neural networks for explicable (mis) classification. *arXiv preprint arXiv:2006.14841*.

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

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*.

Pozanco, A.; Mosca, F.; Zehtabi, P.; Magazzeni, D.; and Kraus, S. 2022. Explaining Preference-driven Schedules: the EXPRES Framework. *arXiv preprint arXiv:2203.08895*.

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.

Rosenfeld, A.; and Richardson, A. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6): 673–705.

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

Sengupta, S.; Chakraborti, T.; and Kambhampati, S. 2018. Ma-radar–a mixed-reality interface for collaborative decision making. *ICAPS UISP*.

Sreedharan, S.; Chakraborti, T.; Muise, C.; and Kambhampati, S. 2020. Expectation-aware planning: A unifying framework for synthesizing and executing self-explaining plans for human-aware planning. *AAAI*.

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

van der Waa, J.; Nieuwburg, E.; Cremers, A.; and Neerincx, M. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291: 103404.

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.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2017. Plan explicability and predictability for robot task planning. In *ICRA*, 1313–1320. IEEE.