Appendix A

In this supplementary material, we present the MILP model of the KP, TAP, WSP and CVRP domains.

Knapsack Problem (KP)

To model the problem we have defined the binary variable $x_{a,i}$ for each agent $a \in A$ and item $i \in I$, where A and I represents a set of agents and items, respectively. These variables will take a value of 1 if item i from agent a is included in the depot.

$$\max \quad \sum_{a \in A, i \in I} x_{a,i} \times \text{UTILITY}(a, i)$$

subject to the following constraint:

$$\sum_{a \in A, i \in I} x_{a,i} \times \text{SPACE}(i) \leq \text{depotCapacity}$$

$$x_{a,i} \in \{0, 1\}$$

The utility of each item $i \in I$ for each argent $a \in A$ is given by $\text{UTILITY}(a, i)$. In our experimental setup utility of each item was between 1 and 5 where 1 represents the lowest and 5 the highest importance. The objective function maximizes the utility of the included items.

Also the problem is subject to a constraint which ensures that the maximum capacity of the depot, depotCapacity, is not exceeded.

Task Allocation Problem (TAP)

To model the problem we have defined the binary variable $x_{a,t}$ for each agent $a \in A$ and task $t \in T$, where A and T represents a set of agents and tasks, respectively. These variables will take a value of 1 if task t is allocated to agent a.

$$\max \quad \sum_{a \in A, t \in T} x_{a,t} \times \text{UTILITY}(a, t)$$

subject to the following constraints:

$$\sum_{t \in T} x_{a,t} \leq \text{WORKLOAD}(a), \quad \forall a \in A$$

$$\sum_{a \in A} x_{a,t} = 1, \quad \forall t \in T$$

$$x_{a,t} \in \{0, 1\}$$

Similar to KP, in this problem the utility of each task $t \in T$ for each argent $a \in A$ is given by $\text{UTILITY}(a, t)$. In our experimental setup utility of each item was between 1 and 10 where 1 represents the lowest and 10 the highest utility. The objective function maximizes the utility of the assigned tasks.

The problem is subject to a set of constraints which ensures that (i) the maximum workload of each agent, $\text{WORKLOAD}(a)$, is respected; and (ii) each task is only allocated to one agent.

Wedding Seating Problem (WSP)

To model the problem, we have introduced set P which is a set of unique pairs of agents p. Each pair $p = (a_i, a_j)$ is consist of two agents $a_i, a_j \in A$, where $i, j \in |A|$ and $i < j$. $\text{AFFINITY}(p)$ represents the affinity value of each pair $p \in P$ (i.e. how much each pair would like to be seated at the same table). Further, we have defined two binary variables $x_{p,t}$ and $y_{a,t}$ for each pair $p \in P$ or each agent $a \in A$ and each table $t \in T$. These variables will take a value of 1 if pair p or agent a is allocated to table t.

$$\max \quad \sum_{p \in P, t \in T} x_{p,t} \times \text{AFFINITY}(p)$$

subject to the following constraints:

$$\sum_{t \in T} x_{p,t} = 1, \quad \forall p \in P$$
$$\sum_{t \in T} y_{a,t} = 1, \quad \forall a \in A$$
$$x_{p,t} \times 2 \leq \sum_{a \in \text{PAIR}(p)} y_{a,t}, \quad \forall p \in P, \quad \forall t \in T$$
$$\sum_{a \in A} y_{a,t} \leq \text{CAPACITY}(t), \quad \forall t \in T$$
$$x_{a,t} \in \{0, 1\}$$

The affinity of each pair of agents $p \in P$ is given by $\text{AFFINITY}(p)$. In our experimental setup affinity of each pair was between 1 and 10 where 1 represents the lowest and 10 the highest affinity. The objective function maximizes the total affinity of pairs allocated to tables.

The problem is subject to a set of constraints. The first and second set of constraints ensure each pair and each agent is assigned to exactly one table. The third set of constraints, makes the connection between two set of binary variables in the problem, $x_{p,t}$ and $y_{a,t}$. $\text{PAIR}(p)$ outputs the set of agents that form pair $p \in P$. The last set of constraints ensures the number of agents assigned to each table respects the capacity of each table, $\text{CAPACITY}(t)$.

Capacitated Vehicle Routing Problem (CVRP)

To model the problem we have defined the binary variable $x_{i,j,v}$ for each point $i \in P$, $j \in P \setminus i$, and $v \in V$, where P is the set of the point representing the points on a map, and V is the set of vehicles that we are planing for. The variable $x_{i,j,v}$ gets value 1 if the vehicle v goes from point i to j.The distance between each two points i and j is given by $\textsc{distance}(i,j)$. It is important to note that the distance between the points are not symmetric.

$$\min \sum_{i \in P, j \in P \setminus i, v \in V} x_{i,j,v} \times \textsc{distance}(i,j)$$

subject to the following constraints:

$$\sum_{j \in P \setminus i} x_{i,j,v} = \sum_{j \in P \setminus i} x_{j,i,v}, \quad \forall i \in P, \forall v \in V$$

$$\sum_{i \in P \setminus p, v \in V} x_{i,b,v} = 1, \quad \forall p \in P \setminus \text{Depot}$$

$$\sum_{i \in P \setminus \text{Depot}} x_{\text{Depot},i,v} = 1, \quad \forall v \in V$$

$$\sum_{j \in P \setminus Depot} x_{\text{Depot},j,v} \leq \textsc{capacity}(v), \forall v \in V$$

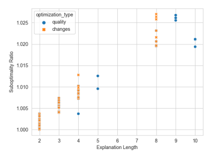
$$\sum_{i \in \textsc{set}, j \in P \setminus \textsc{set}, v \in V} x_{i,j,v} \geq 1, \quad \forall \textsc{set} \in \textsc{subsets}(P \setminus \text{Depot})$$

$$x_{i,j,v} \in \{0,1\}$$

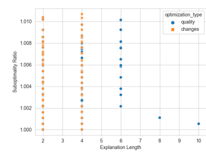
The objective function minimizes the total traveled distance by all the vehicles. The first constraint ensures that all vehicles leave the points they visit. The second constraint ensures that each point is visited exactly once by any vehicle. The third constraints ensure all the vehicles start from the depot. The return of vehicles to the depot is ensured by the conjunction of the first and third constraints. The fourth constraint ensures that the capacity of the vehicle, i.e., the number of points a vehicle can visit is satisfied. Finally, the fifth constraint prevents the existence of subtours in the returned routes.This is done by pre-computing all the subsets composed by the points in $P \setminus$ Depot.

Appendix B

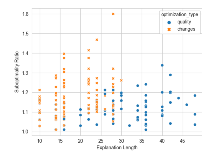
Solution Quality vs Explanation Length trade-off in the four domains.



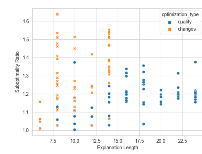
(a) KP



(b) TAP

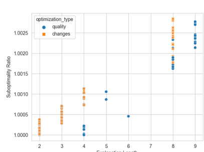


(c) WSP

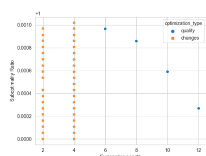


(d) CVRP

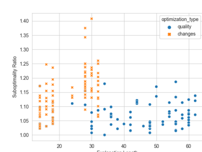
Figure 1: Trade-off between Explanation Length and Suboptimality Ratio when solving the HMAOP with Q-CMAoE (blue points) or C-CMAoE (orange crosses). Problems with 100 agents in KP, 100 tasks in TAP, 10 agents in WSP, and 10 points in CVRP.



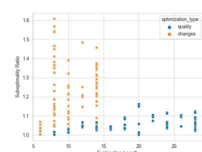
(a) KP



(b) TAP



(c) WSP



(d) CVRP

Figure 2: Trade-off between Explanation Length and Suboptimality Ratio when solving the HMAOP with Q-CMAoE (blue points) or C-CMAoE (orange crosses). Problems with 1000 agents in KP, 1000 tasks in TAP, 12 agents in WSP, and 12 points in CVRP.

Appendix C

Knapsack scenario used in our user study. The below scenario is shown in raw format and without any interactive component, while the user study was run in an interactive platform. Please refer to the main body of the paper for details on how experiments were run, i.e., explanations shown per user, how the scenario was introduced to the participants etc.

Scenario. Consider we have 7 people (Tal, Noam, Dagan, Bar, Gefen, Aviv and Ziv) that want to share a storage with space capacity of 55. Each person has the following items: bed, sofa, table, chair, lamp, books, computer, clothes, fridge, and fan. However, each item has a different utility for each person. The following table specifies the space required to include each item.

Item	Bed	Sofa	Table	Chair	Lamp	Books	Computer	Clothes	Fridge	Fan
Space	25	12	8	4	2	2	2	3	4	2

Table 1: Space required to include each item.

Optimal solution. We computed the optimal solution for this problem. The optimal solution would allow only the following items from each person to be in the storage:

- Tal: computer, fan
- Noam: chair, lamp, computer
- Dagan: lamp
- Bar: lamp, books, computer, fan
- Gefen: chair, computer
- Aviv: lamp, books, computer, clothes
- Ziv: table, lamp, books, fridge, fan

Question and Explanation. Imagine you are Tal and the utility of each item for you is shown below, where the higher the utility level indicates the higher importance of the item.

Item	Bed	Sofa	Table	Chair	Lamp	Books	Computer	Clothes	Fridge	Fan
Utility	4	2	2	1	1	1	1	1	1	1

Table 2: Tal's utility assigned to each item.

Considering that you are Tal, please mark the most accurate statement.

I’m dissatisfied with the allocation	I’m somewhat dissatisfied with the allocation	I’m neither satisfied nor dissatisfied with the allocation	I’m somewhat satisfied with the allocation	I’m satisfied with the allocation
○	○	○	○	○

Please mark to what extent do you agree with the following statement:

I would like to make a complaint about my allocation.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
◯	◯	◯	◯	◯

We would like to present you with an explanation regarding the allocation that you (Tal) have received.

Considering that you (Tal) are dissatisfied with the allocation, you have asked: ”Why is my bed not included in the optimal solution?”

Provided explanation.

Here is your explanation:

- Sorry, this is what the algorithm generated
- Total utility would decrease
- Total utility would decrease by 13 based on the following table:

Item	Tal	Dagan	Gefen	Aviv	Ziv
Removed items (utility)	Computer, Fan (2)	Lamp (1)	Chair (4)	Clothes(2)	Fridge, Table (8)
Added items (utility)	Bed (4)	-	-	-	-

Please mark the most accurate statement regarding your satisfaction with the explanation.

I'm dissatisfied with the explanation	I'm somewhat dissatisfied with the explanation	I'm neither satisfied nor dissatisfied with the explanation	I'm somewhat satisfied with the explanation	I'm satisfied with the explanation
◯	◯	◯	◯	◯

Considering the provided explanation, we would like to ask again, please mark the most accurate statement.

I'm dissatisfied with the allocation	I'm somewhat dissatisfied with the allocation	I'm neither satisfied nor dissatisfied with the allocation	I'm somewhat satisfied with the allocation	I'm satisfied with the allocation
○	○	○	○	○

Please mark to what extent do you agree with the following statement:

I would like to make a complaint about my allocation.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
◯	◯	◯	◯	◯

Finally, we would like to understand your satisfaction from the explanations provided. Could you please mark to what extend you agree with the following statements.

*From the explanation, I **understand** how the tool works.*

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
◯	◯	◯	◯	◯

*This explanation of how the tool works is **satisfying.***

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
◯	◯	◯	◯	◯

*This explanation of how the tool works has **sufficient detail.***

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
◯	◯	◯	◯	◯

*This explanation of how the tool works seems **complete.***

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
◯	◯	◯	◯	◯

*This explanation of how the tool works **tells me how to use it.***

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
○	○	○	○	○

*This explanation of how the tool works is **useful to my goals.***

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
○	○	○	○	○

*This explanation of the tool shows me how **accurate** the tool is.*

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
○	○	○	○	○

*This explanation lets me judge when I should **trust and not trust** the tool.*

Strongly disagree	Disagree	Neutral	Agree	Strongly agree
○	○	○	○	○

Appendix D

Based on ANOVA analysis, in all domains, we found a *main effect* of satisfaction with the solution after an explanation was presented. This means there was a change in the mean of satisfaction from the solutions before and after presenting an explanation. The main effect was statistically significant in all domains; KP ($F(1) = 41.26$, $p < 0.01$), WSP ($F(1) = 95.70$, $p < 0.01$), CVRP ($F(1) = 12.16$, $p < 0.01$) and TAP ($F(1) = 15.49$, $p < 0.01$), where the p-value was less than 0.05. An *interaction effect* between the change in satisfaction with the solution and the type of explanation was found in all domains. The interaction effect represents whether the changes in satisfaction with the solution were different based on different types of explanations. The F-scores and p-values were: KP ($F(2) = 10.48$, $p < 0.01$), WSP ($F(2) = 11.29$, $p < 0.01$), CVRP ($F(2) = 6.89$, $p < 0.01$) and TAP ($F(2) = 2.62$, $p = 0.08$). In all domains, the p-value indicated a strong significance, except in the TAP domain which was close to significant. Similarly, in all domains, we found a *main effect* of decrease on the desire to complain following an explanation. The main effect was statistically significant in all four domains: KP ($F(1) = 27.45$, $p < 0.01$), WSP (87.50, $p < 0.01$), CVRP ($F(1) = 27.02$, $p < 0.01$) and TAP ($F(1) = 24.66$, $p < 0.01$), where the p-value was less than 0.05. Also, an *interaction effect* was found in all domains between the change in desire to make a complaint and the type of explanation . The F-scores and p-values for the four domains were: KP ($F(2) = 5.3$, $p < 0.01$), WSP ($F(2) = 11.29$, $p < 0.01$), CVRP ($F(2) = 9.3$, $p < 0.07$) and TAP ($F(2) = 2.71$, $p = 0.07$). In all domains, the p-value indicated a strong significance except in the TAP domain, which was close to significant.

Appendix E

Based on ANOVA analysis, in the analysis of satisfaction with the solution for the KP domain, we found a *main effect* for the change in satisfaction with the solution following the explanation ($F(1) = 28.8, p < 0.01$). No effect was found for the type of explanation ($F(1) = 1.25, p = 0.2$). Also, no *interaction effect* was found between the change in satisfaction with the solution and the type of explanation ($F(1) = 1.64, p = 0.2$). In the KP domain, a *main effect* was found for the change in the desire to complain following the explanation ($F(1) = 42.49, p < 0.01$) and no effect was found for the type of explanation ($F(1) = 1.7, p = 0.2$). However, an *interaction effec*t was found between the desire to complain and the type of explanation $F(1) = 7.2, p = 0.01$).

Similarly, in the WAP domain no *main effects* were found for satisfaction with the solution following the explanation ($F(1) = 0.37, p = 0.5$) and the type of explanation ($F(1) = 0.13, p = 0.7$). Furthermore, no *interaction effect* was found between the change in satisfaction with the solution and the type of explanation ($F(1) = 0.01, p = 0.9$). In the analysis of the desire to complain in the WAP domain, we found *main effects* for the change in the desire to complain following the explanation ($F(1) = 8.2, p < 0.01$). No effect was found for the type of explanation ($F(1) = 0.55, p = 0.46$). In addition, no *interaction effect* was found between the desire to complain and the type of explanation ($F(2) = 1.75, p < 0.2$).

Appendix F

- **S1:** From the explanations, I **understand** how the optimization model works.
- **S2:** The explanations are **satisfying**.
- **S3:** The explanations are sufficiently **detailed**.
- **S4:** The explanations are sufficiently **complete**, that is they provide me with all the necessary information to decide whether to complain.
- **S5:** The explanations are **actionable**, that is, they help me know how to answer the questions.
- **S6:** The explanations let me know how **reliable** the optimization model is.
- **S7:** The explanations let me know how **trustworthy** the optimization model is.

Statement	Explanation	Domains			
		KP		WAP	
		N	μ (std)	N	μ (std)
S1	CMAoE	19	3.95 (0.78)	22	3.36 (1.29)
	KORIKOV21	21	2.67 (1.2)	20	3.1 (1.11)
	Total	40	3.28 (1.2)	42	3.24 (1.21)
S2	CMAoE	19	3.58 (1.17)	22	2.73 (1.08)
	KORIKOV21	21	2.38 (1.2)	20	2.85 (0.99)
	Total	40	2.95 (1.32)	42	2.79 (1.023
S3	CMAoE	19	3.47 (1.12)	22	3.36 (1.18)
	KORIKOV21	21	2.29 (1.31)	20	2.65 (1.27)
	Total	40	2.85 (1.35)	42	3.02 (1.26)
S4	CMAoE	19	3.58 (1.22)	22	3.18 (1.18)
	KORIKOV21	21	2.48 (1.47)	20	2.75 (1.21)
	Total	40	3 (1.45)	42	2.98 (1.2)
S5	CMAoE	19	3.74 (0.93)	22	3.27 (1.12)
	KORIKOV21	21	2.86 (1.39)	20	3.25 (1.37)
	Total	40	3.28 (1.26)	42	3.26 (1.23)
S6	CMAoE	19	3.16 (1.17)	22	2.73 (1.16)
	KORIKOV21	21	2.05 (0.97)	20	2.85 (1.18)
	Total	40	2.58 (1.2)	42	2.79 (1.16)
S7	CMAoE	19	3.32 (1.06)	22	2.86 (0.99)
	KORIKOV21	21	2.24 (1.22)	20	2.5 (1.23)
	Total	40	2.75 (1.14)	42	2.69 (1.12)

Table 16: User's satisfaction (mean (std)) for good metrics' staements with each explanation in the two domains. N represents the number of participants per session.