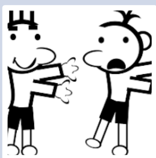
AI for Explaining Decisions in Multi-Agent Environments

Sarit Kraus
Bar-Ilan University
Israel

Bar-Ilan University
אוניברסיטת בר־אילן
J.P.Morgan
TAILOR

Decisions in Multi-agent Environments (including humans & robots)



Agents have possibly conflicting preferences.

The AI system (e.g., COMSOC algo) should balance between these preferences

A decision may make some people unhappy.

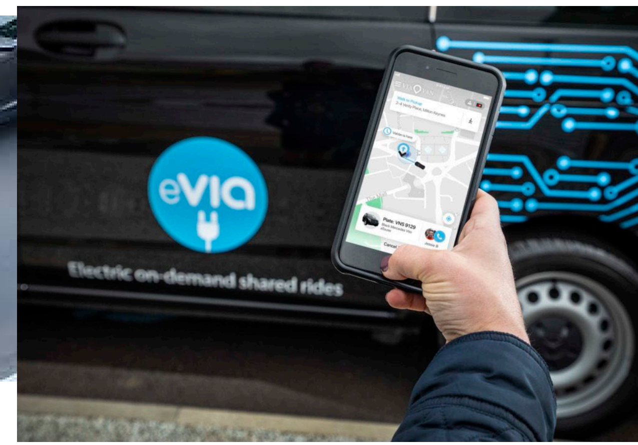
Shared workspaces: resource allocation

Matching


Let's Carpool!
SHARE A CAR


1-800-VAN-RIDE


CARPOOL
by waze


eVIA
Electric on-demand shared rides

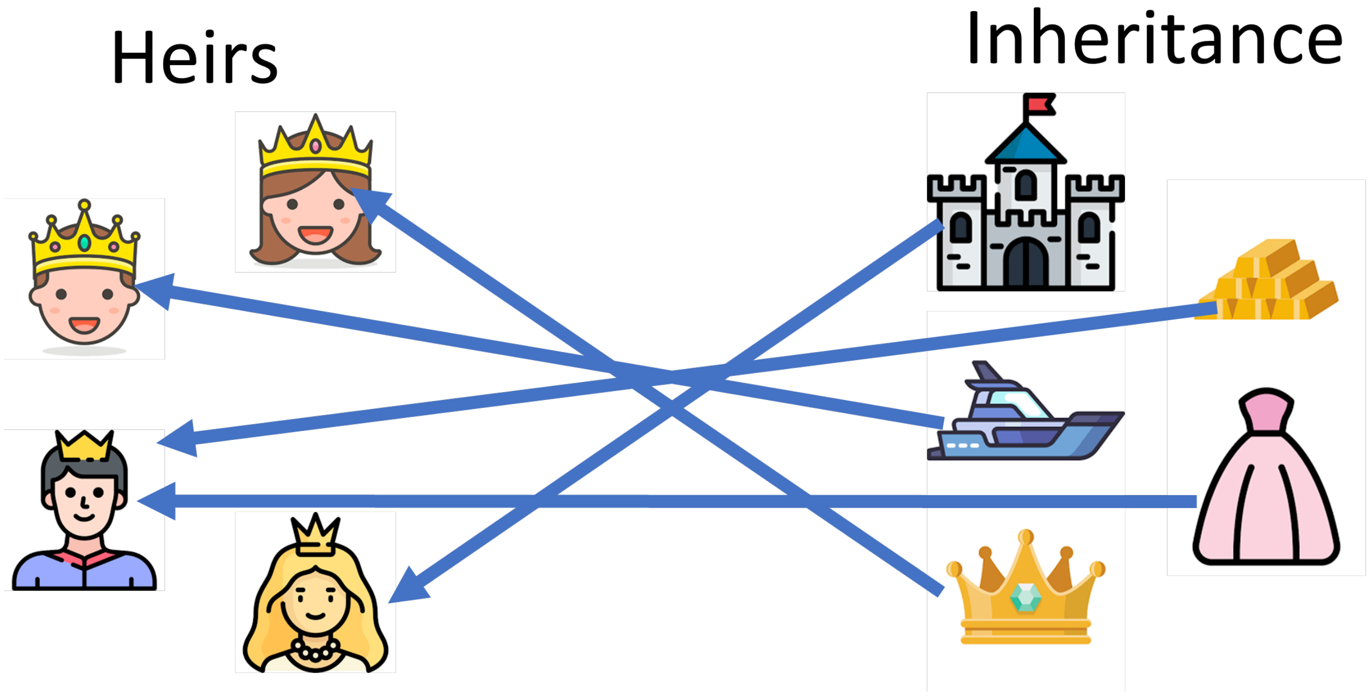

POOL IT


bubble Dan

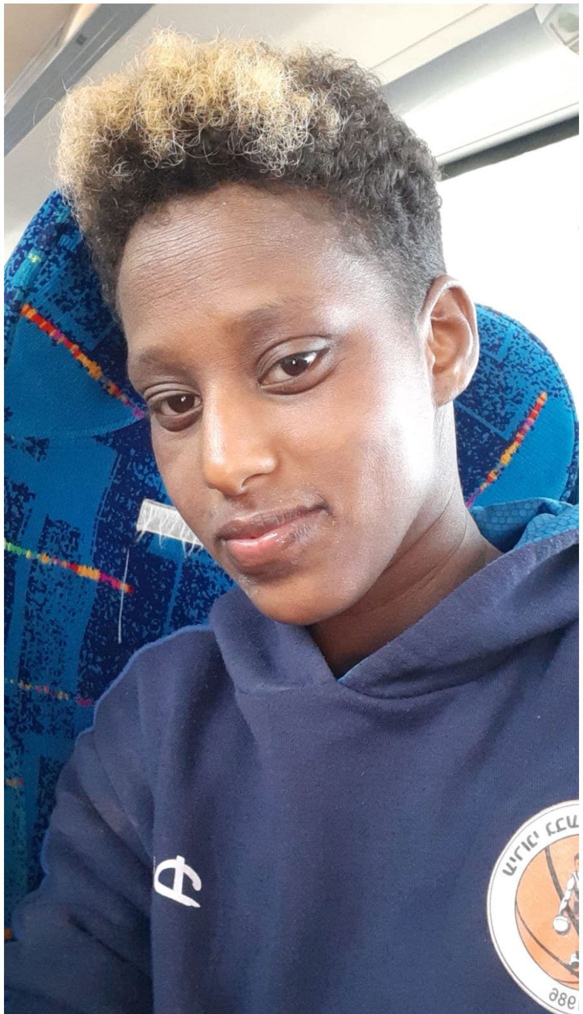
Scheduling

	8am-9am	9am-10am	10am-11am	11am-12pm	12pm-1pm	1pm-2pm	2pm-3pm	3pm-4pm
Bob					*	*	*	*
Alice	*	*	*	*				

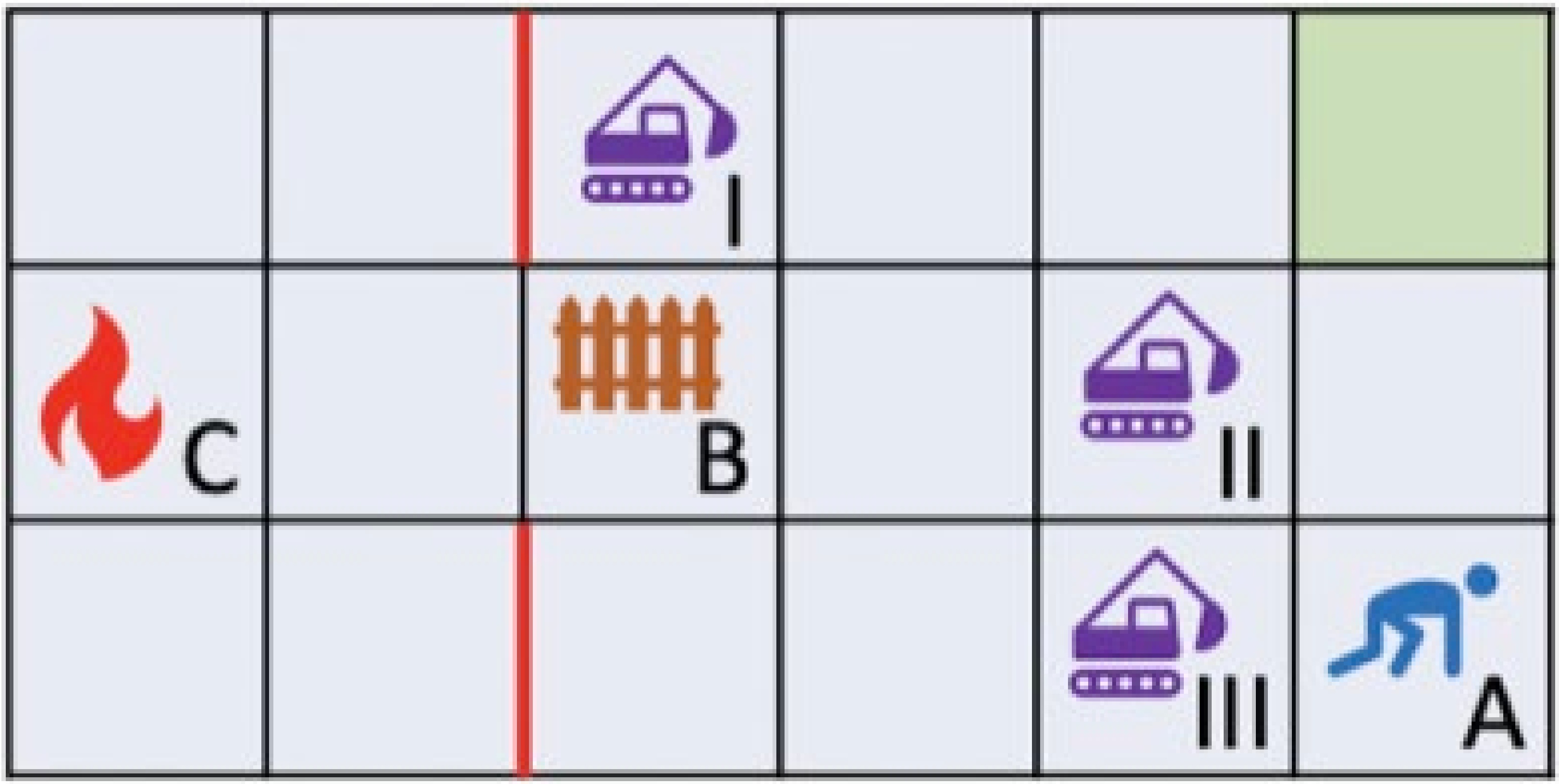
Fair division of indivisible goods



Group formation:
Dividing students to classes



Multi-agent planning: search and rescue



Explainable decisions in Multi-Agent Environments (xMASE)

- Providing explanations about the system's decision:
  - increases people's satisfaction
  - maintains acceptability of the AI system
  - satisfies regulation; EU General Data Protection Regulation: “meaningful information of the logic involved” for automated decisions.

XAI vs xMASE

- **XAI**
  - explains to a user a decision made by an AI blackbox system.
  - AI blackbox maximizes a well-agreed upon function
  - Main goal: increasing users' trust in the black-box AI system
- **xMASE**
  - the maximization function is not clear to the (human) agents due to unknown others' preferences
  - Goal: increase user satisfaction, taking into account properties such as fairness, envy and privacy.

xMASE

- Need to refer to
  - technical reasons that led to the decision (XAI)
  - preferences of the agents that were involved
  - fairness
- Challenges:
  - What to reveal from other agents' preferences?
    - privacy of other agents
    - how these preferences led to the final decision.
  - The influence of the explanation on user satisfaction changes from one user to the next;
    - personalized explanations

Evaluation of Explanations: people

- Many algorithms that provide explanations on AI systems take an engineering approach, which does not involve running experiments with people.

xMASE

- Explaining Preference-Driven Schedules (ICAPS 2022)
- Justifying Social-Choice Mechanism Outcome for Improving Participant Satisfaction (AAMAS 2022)
- Towards Policy Explanations for Multi-Agent Reinforcement Learning, (IJCAI 2022)
- Explainable Multi-Agent Reinforcement Learning for Temporal Queries (IJCAI 2023)

- Why do explanations help? -- they are cheap talk?
- Can we build formal models that will include communication, for example, on fairness, and will influence the agents' strategies.

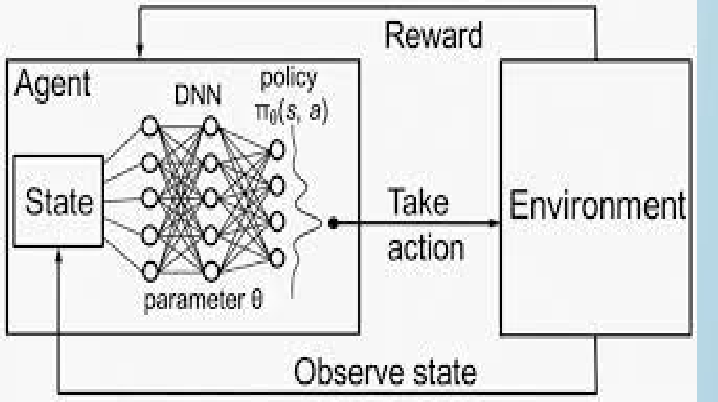
Explainable Multi-Agent Reinforcement Learning For Temporal Queries

- Kayla Boggess[1], Sarit Kraus[2], and Lu Feng[1]
- [1]University of Virginia
- [2]Bar-Ilan University
- {kjb5we, lu.feng}@virginia.edu, sarit@cs.biu.ac.il

IJCAI 2023, August 19-25

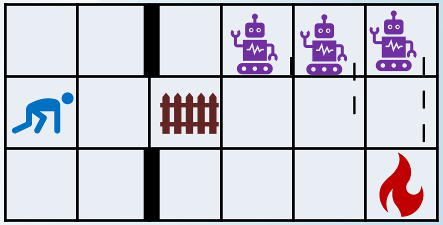
Motivation: explaining MARL black box policy

- Increasing deployment of MARL systems in society
  - Systems may be too complex for users to understand
- Why do we need to explain MARL?
  - Improve system transparency
  - Higher understandability
  - **Increase user satisfaction**
  - Increase agent trust
  - **Better human-agent cooperation**



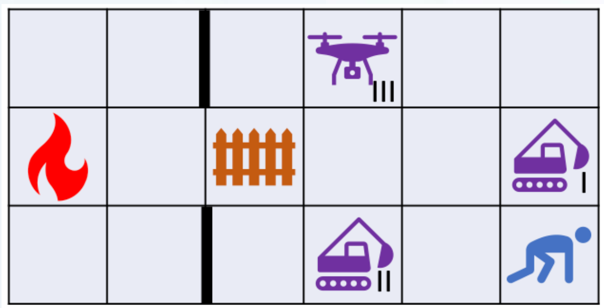
Temporal Queries

- Need to address query types about agent behavior
  - Contrastive (IJCAI22)
    - Why event *p* and not event *q*?
    - *Why agent 1 and agent 2 remove obstacle and not fight fire at time 1?*
  - Temporal (IJCAI23)
    - Why not *task 1* and then *task 2?*
    - Why not agent 1 and agent 2 *remove obstacle* and then agent 3 *fight fire*?



- Generate policy-level contrastive explanations for multi-agent reinforcement learning
  - Explain if an alternate plan is feasible under a given policy

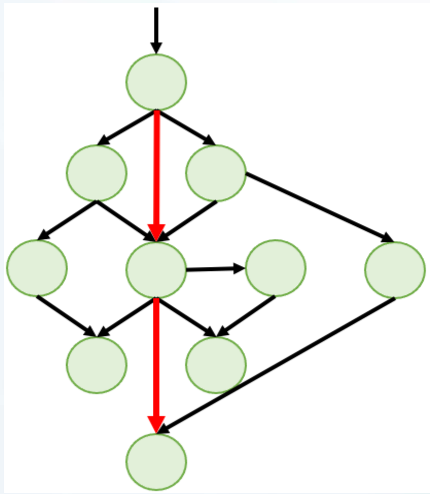
Policy Abstraction

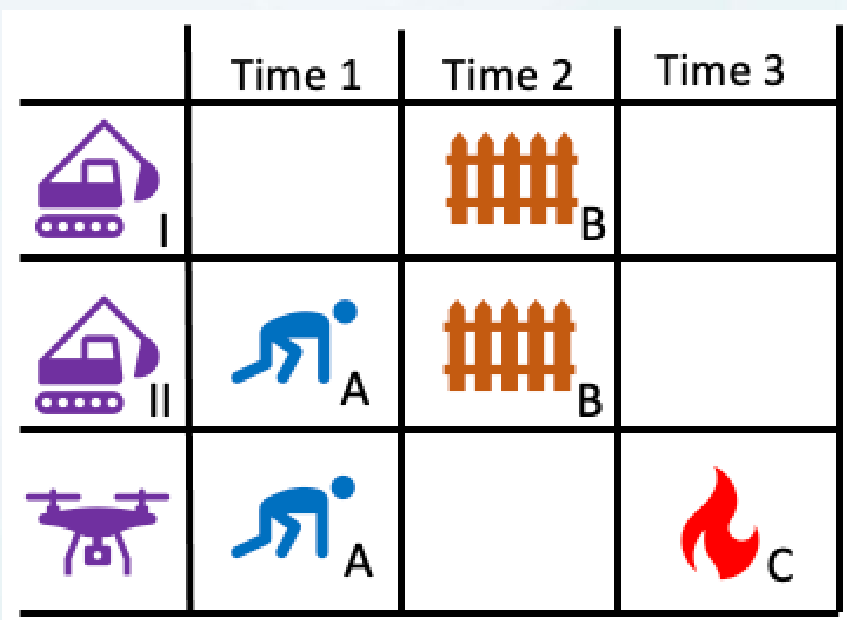


Victim_Detected
Obstacle_Not_Detected
Fire_Not_Detected
Victim_Not_Complete
Obstacle_Not_Complete
Fire_Not_Complete

- Train joint policy for *N* agents
- Generate abstract features for new state space containing adequate information for explanation
- Convert each training sample to corresponding abstract state
- Compute transition probability via frequency counting

Policy Summarization





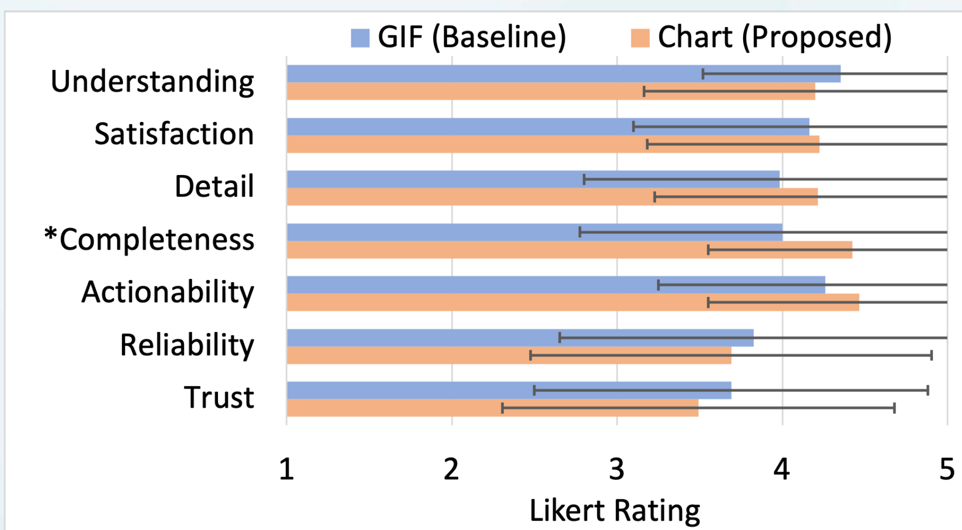
- Search policy graph for most probable path
- Extract agent cooperation and task sequences (If an agent satisfies a task, assign task to all agents involved.)
- Generate chart to show to user

Hypotheses

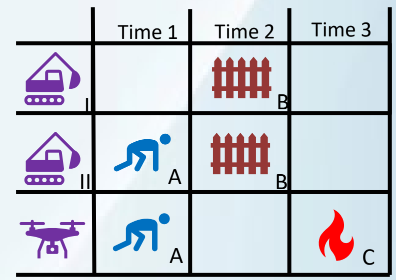
- H1: Chart-based summarizations lead to better user performance than GIF-based.
- H2: Chart-based summarizations yield higher user ratings on explanation goodness metrics than GIF-based.

User Study: Policy Summarization (116 subjects)

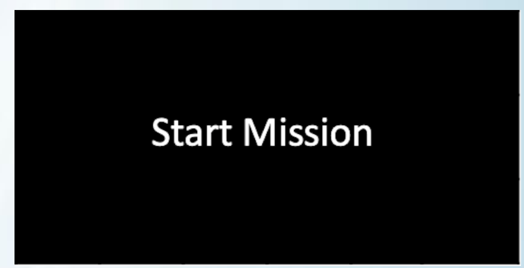
- **User Performance**
  - Proposed: M=1.8 out of 2, SD=0.6
  - Baseline: M=0.9 out of 2, SD=0.4
- **Explanation Goodness**



Proposed:



Baseline:



Sample Question: Which robots are required to save victim A?

User Query

- An alternate plan presented by the user
- States task order and agent cooperation requirements
- Can be feasible or infeasible
- Can be full (all tasks) or partial (some tasks)
  - Unmentioned tasks can be completed in any order
- Can contain any of the tasks present in the environment

Original Plan

	Task 1	**Task 2**	**Task 3**
Robot I	-	Obstacle	Victim
Robot II	Fire	Obstacle	-
Robot III	Fire	-	Victim

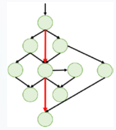
User Query

	Task 1	**Task 2**	**Task 3**
Robot I	Obstacle	Victim	-
Robot II	Obstacle	-	Fire
Robot III	-	Victim	Fire

Construct a policy abstraction MMDP given π

User Query

	Task 1	Task 2	Task 3
Robot I	Obstacle	Victim	-
Robot II	Obstacle	-	Fire
Robot III	-	-	Fire

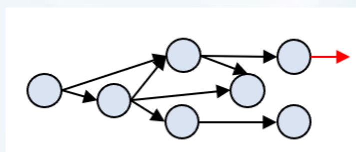


$$\varphi = \mathrm{P}_{>0}[\Diamond(\tau_1 \wedge \Diamond(\tau_2 \wedge \Diamond \cdots))]$$ **PCTL***

1 Query Checking with Temporal Logic

Feasible Query

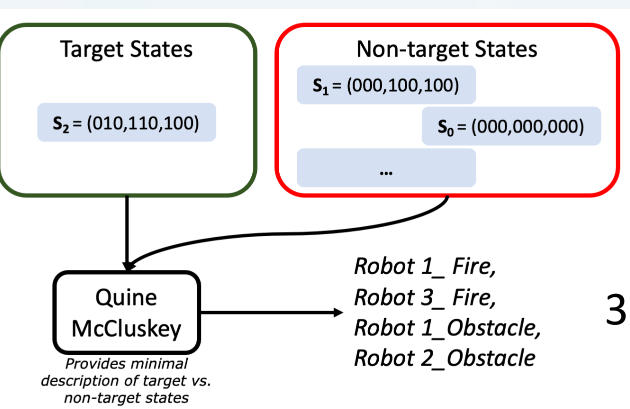
"Your plan is feasible."



2 Guided Rollout

Feasible Query

"Your plan is feasible."



3 Explanation Generation

Infeasible Query

"The robots cannot rescue the victim because Robot I needs Robot III to help rescue the victim."

	Task 1	Task 2
Robot I	-	-
Robot II	Fire	Obstacle
Robot III	Fire	-

Guided Rollout

- Guided rollout procedure to sample more of the MARL agents' behaviors and update the MMDP with new samples.
- The search is motivated by the query: Apply a U-value to each state measures how close the state is to the user's query
- Check PCTL* formula for feasibility

Explanation Generation

- Find highest U-value in policy abstraction
  - U+1 is failed task causing infeasibility

- Find target and non-target states
  - Target – States where task is completed
  - Non-target – All other possible states
  - No target states means task is impossible in observed states
- Run Quine-McCluskey to find the minimal number of terms that are different between the target and non target state with the highest U.
- Generate an explanation using a natural language template (GPT)

Case Study				MMDP		Feasible	Infeasible	
Domain	Agents	Tasks	|Query|	$\|\mathcal{S}\|$	$\|\mathcal{T}\|$	Time (s)	# Failed Tasks	Time (s)
SR	3	3	3	28	127	0.8	1	2.2
	4	4	4	163	674	1.5	2	5.3
	5	5	5	445	1,504	24.4	3	89.8
LBF	3	3	3	67	344	0.9	1	2.9
	4	4	4	211	781	2.1	2	7.6
	5	5	5	152	454	4.5	3	20.5
RWARE	2	4	3	98	268	0.8	1	15.5
	3	6	5	442	1,260	3.7	2	42.2
	4	8	8	1,089	2,751	21.7	3	85.2
PLATE	5	3	3	87	181	0.8	1	3.0
	7	4	4	85	175	0.9	2	25.7
	9	5	5	132	266	1.4	3	126.8

Table 1: Experimental results on four benchmark MARL domains.

- Infeasible queries take significantly more time due to guided rollout and explanation generation
- The time to generate explanations scales with number of failed tasks

Hypotheses

- H1: Explanations generated by our proposed approach enable participants to answer more questions correctly than the baseline explanations.
- H2: Explanations generated by our proposed approach lead to higher ratings on explanation goodness metrics than the baseline explanations.

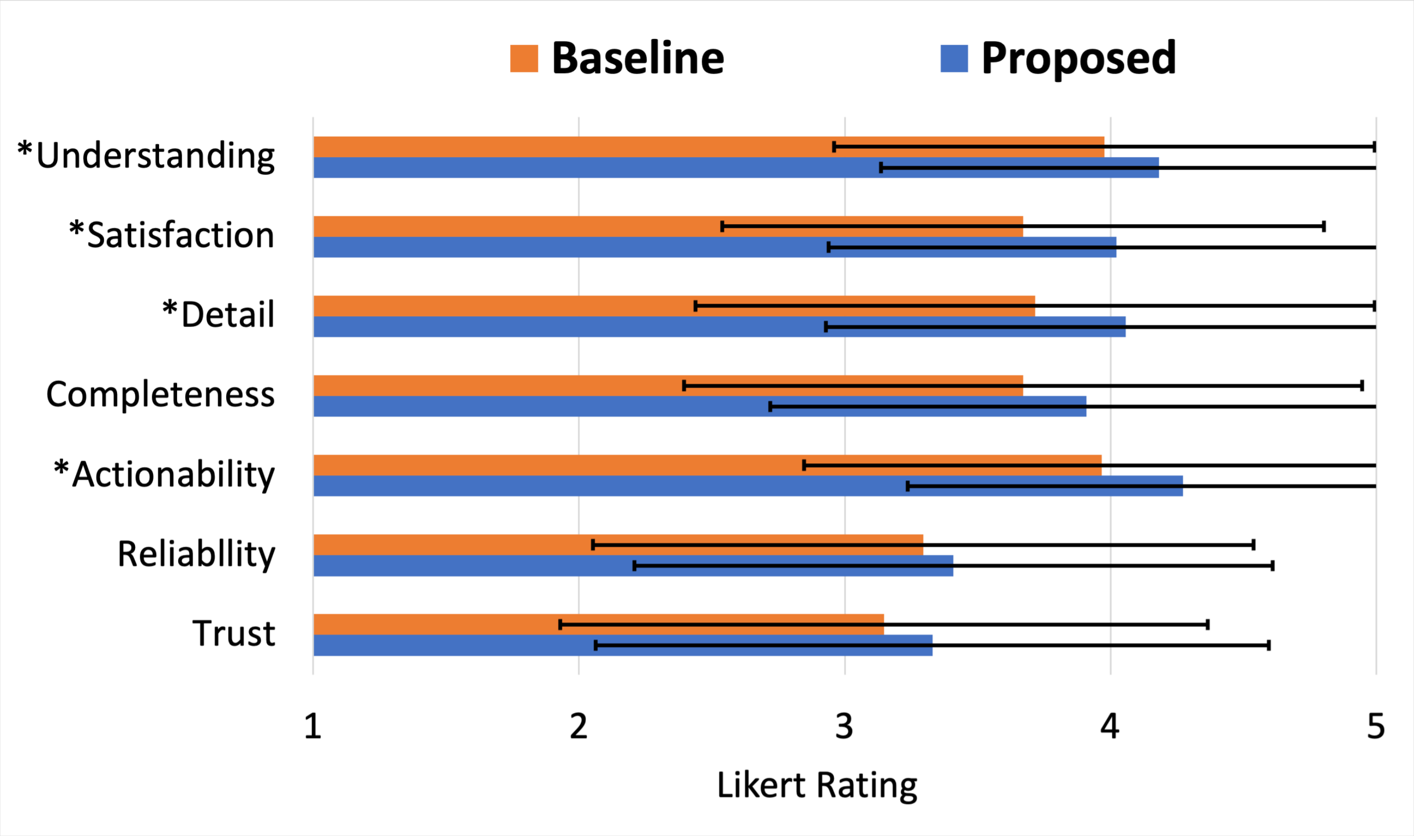
Study Design

- **Goal:** Given an original plan and alternate plan, predict if new plan is feasible based on a provided explanation for the alternate plan.
- **Baseline:** "Bridging the Gap" by Sreedharan et. al., 2022.
- **Background:** 88 Participants
  - Bonus payment for correctly answered questions
  - Demonstrations, attention checks implemented, time to complete tracked
  - 2 trials (proposed, baseline) of 4 questions
- **Hypotheses:**
  - Enable participants to answer more questions
  - Leads to higher ratings on explanations goodness metrics (adapted from Hoffman et al., 2018)

The study was approved by UVA Institutional Review Boards (IRB)

Results: 88 Participants

- **User Performance**
  - Proposed: M=3.1 out of 4, SD=1.0
  - Baseline: M=0.6 out of 4, SD=0.8
- **Explanation Goodness**



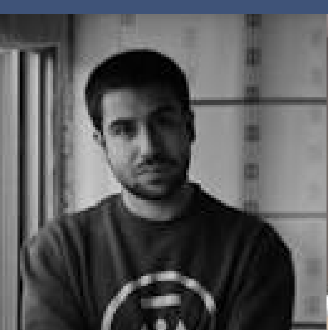
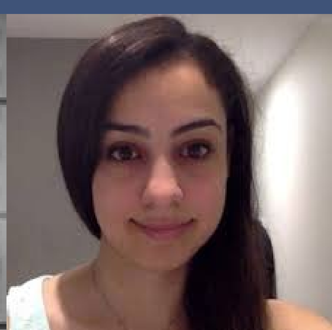
Conclusions

- Developed a method to generate explanations to answer temporal user queries for multi-agent reinforcement learning
- Applied method to four MARL benchmarks to show effectiveness and scalability
- Conducted user study to evaluate quality of explanations

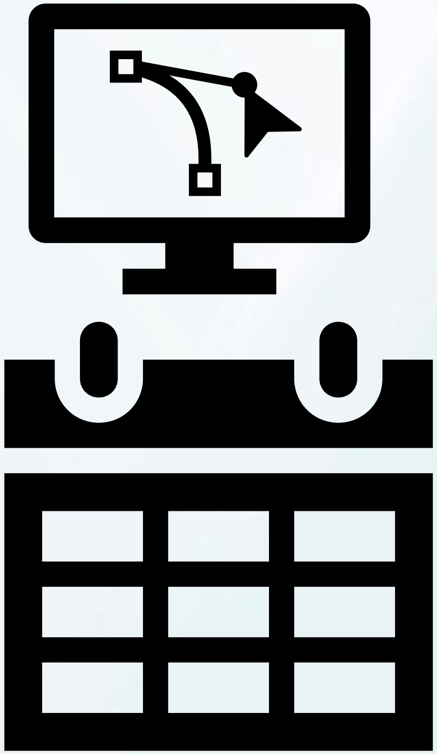
Contrastive Explanations of Multi-agent Optimization Solutions

Alberto Pozanco, Parisa Zehtabi, Ayala Bloch,
Daniel Borrajo Sarit Kraus

Bar-Ilan University
J.P.Morgan AI RESEARCH
ARIEL UNIVERSITY

Explaining general Optimization Problems

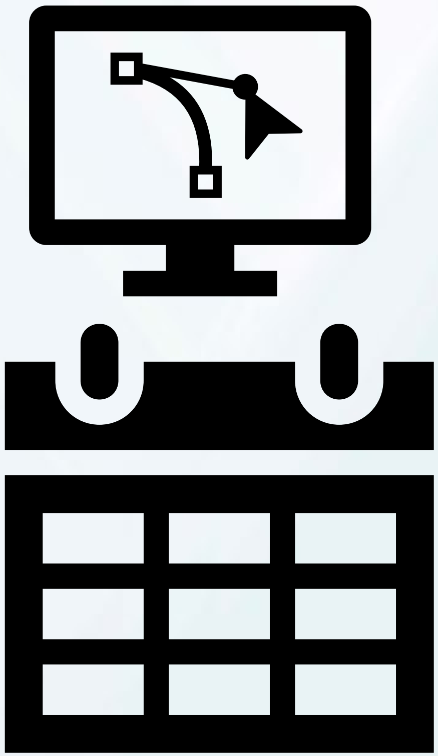


General cost function

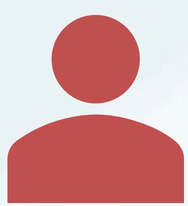
Constraints

Meaningful variables names
Specification of relevant variables

Ongoing work: explaining general Optimization Problems

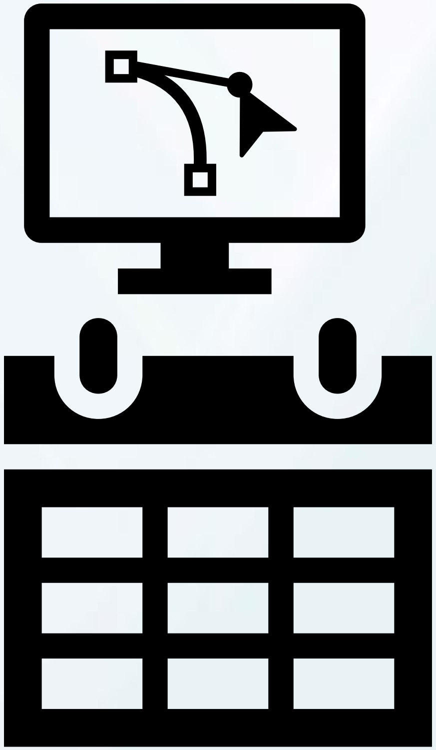


General cost function

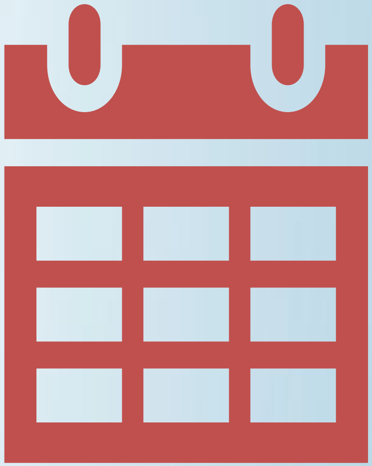


Why not a solution with property X?

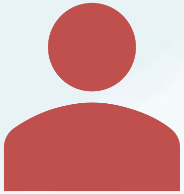
Explaining General Optimization Problems



Hypothetical optimization problem with constraints X



Explain diff between both solutions

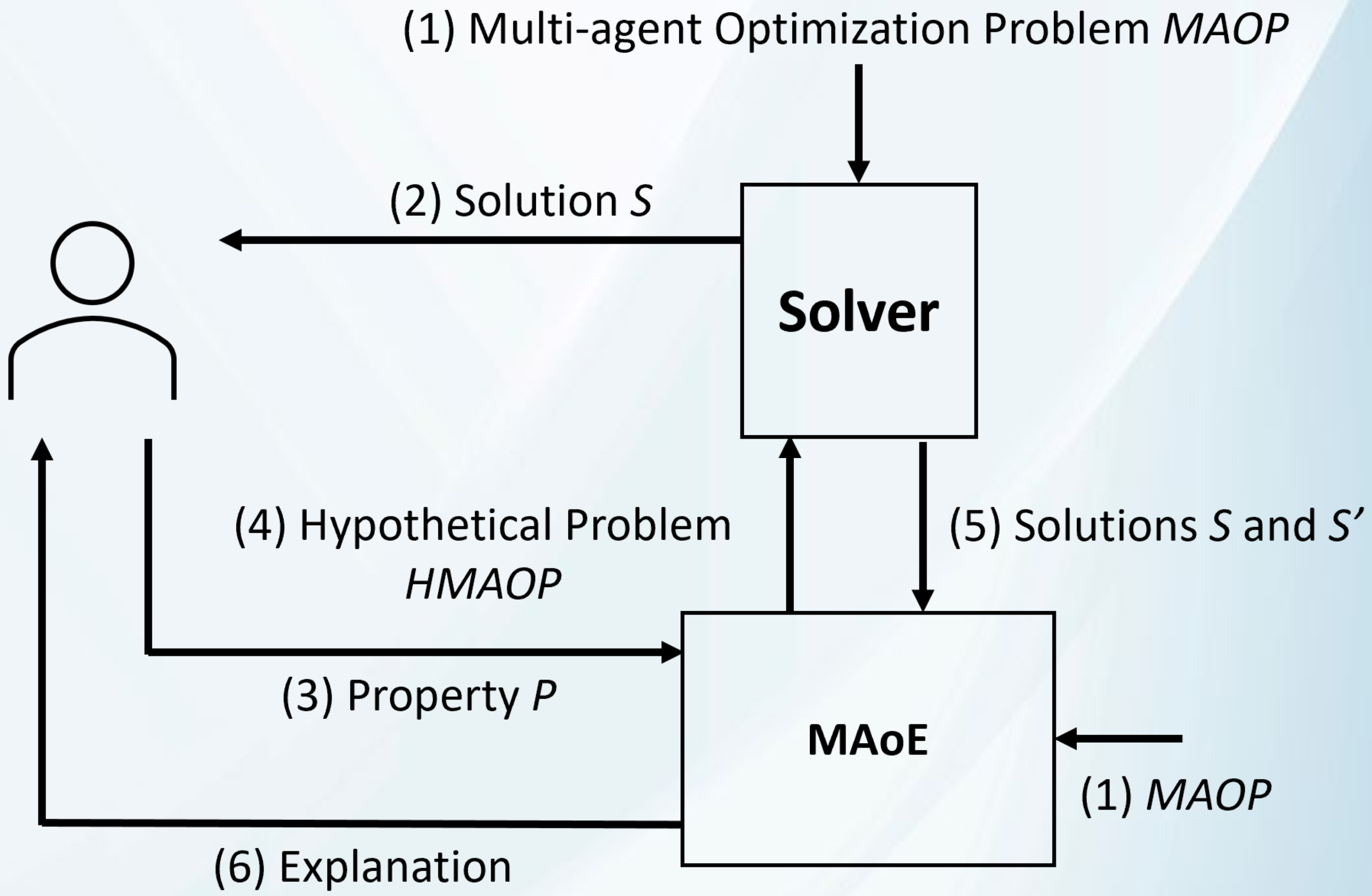


Why not a solution with property X?

Explaining general Optimization Problems

Hypothetical optimization problem with constraint X:

- Optimal among the most similar to the original solution
- Most similar among the optimal solutions

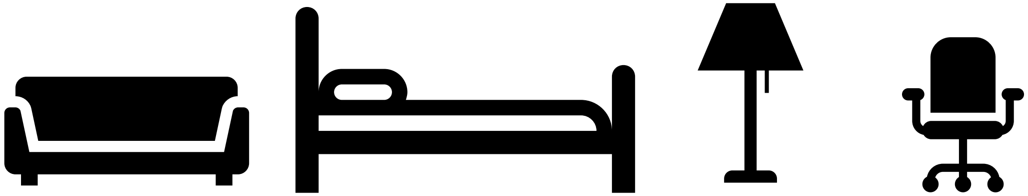


(1) Multi-agent Optimization Problem *MAOP*
(2) Solution *S*
Solver
(4) Hypothetical Problem *HMAOP*
(5) Solutions *S* and *S’*
(3) Property *P*
MAoE
(1) *MAOP*
(6) Explanation

Knapsack

<u>Objective</u>: Maximize total utility in the container while satisfiying its capacity

Container Capacity C



Space: A
Utility: W

Space: B
Utility: X

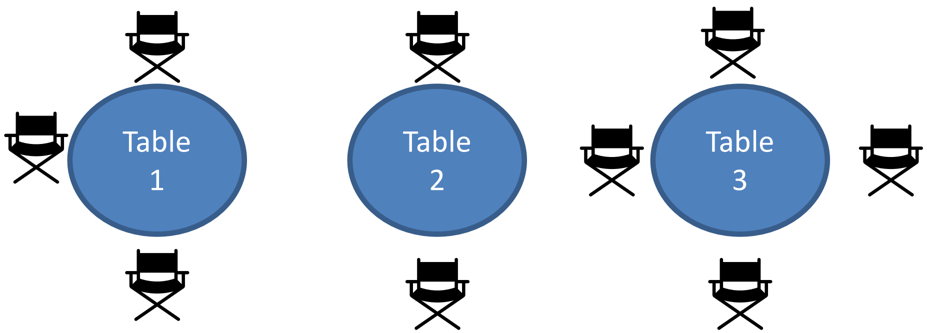
Space: C
Utility: Y

Space: D
Utility: Z

- Each agent has a set of items to include
- Each item has a different utility for each agent

Wedding Seating

<u>Objective</u>: Maximize total friendship while satisfying tables' constraints



- Each pair of agents has a friendship level

- Each table can fit a different number of people

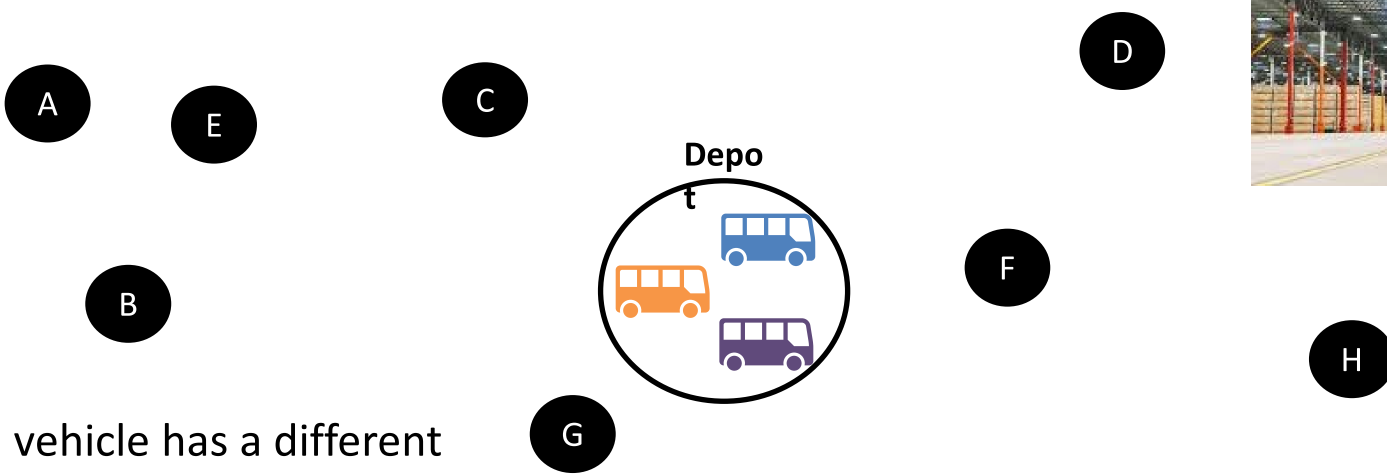
Task Allocation

<u>Objective</u>: The goal is to maximize the total utility.

- Each agent has a different utility for each task

Vehicle Routing Problem (VRP)

<u>Objective</u>: Minimize total travelled distance while satisfying vehicles' constraints



- Each vehicle has a different capacity, i.e., number of points it can visit
- All vehicles must start and finish the routes in the Depot

Hypotheses

- H1: Explanations improve humans' satisfaction with the decisions made by the AI system.
- H2: Explanations reduce humans' desire to complain about the decisions made by the AI system.
- H3: Humans prefer more detailed explanations.

User study procedures

- Explain the setting
- Provide a solution
- Ask for satisfiability from the solution and desire to complain
- Provide explanation
- Ask for satisfiability from the solution and desire to complain
- Baseline: ``Sorry, this is what the algorithm generated''

User Study

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
○	○	○	○	○

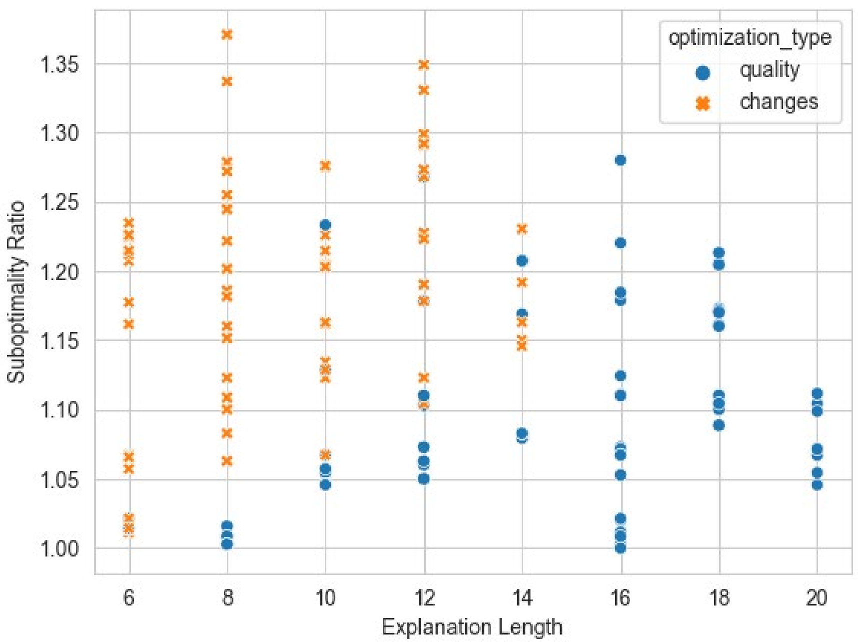
Wedding Seating Explanations

Placebo: ``Sorry, this is what the algorithm generated''
Short: Total friendship will decrease
Detailed: Total friendship would decrease by 10 based on the following table:

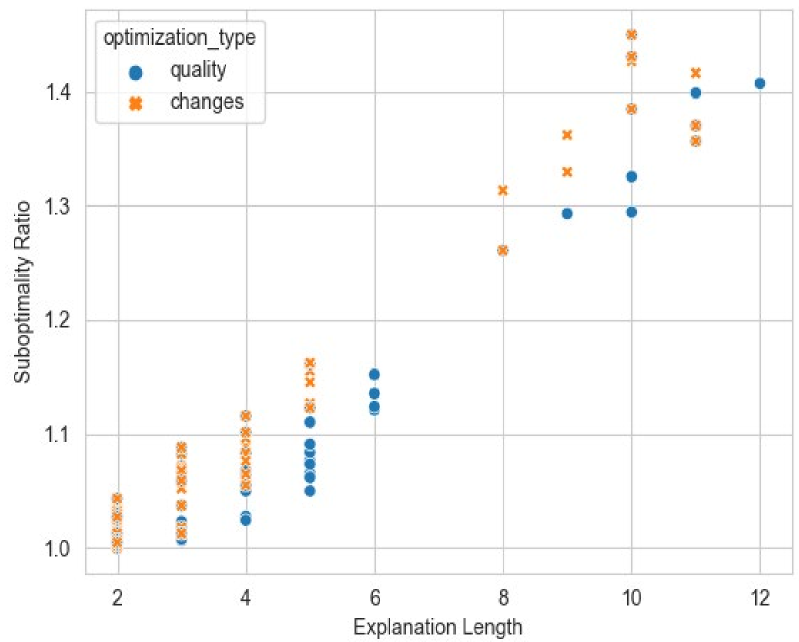
	Ziv	Gabi	Gefen	Lee	Agam	Aviv	Bar	Noam	Tal	Dagan
Not seated together anymore	Noam (6)	Lee (5)	Noam (7)	Aviv, Gabi, Agam, Bar (21)	Lee (7)	Lee (1)	Lee (8)	Ziv, Gefen (13)	Dagan (1)	Tal (1)
Seated together Now	Lee (4)	Dagan (6)	Lee (1)	Gefen, Ziv (5)	Dagan (2)	Dagan (1)	Dagan (7)	Tal (9)	Noam (9)	Gabi, Aviv, Agam, Bar (16)

Suboptimality vs Explanation Length

Vehicle Routing Problem

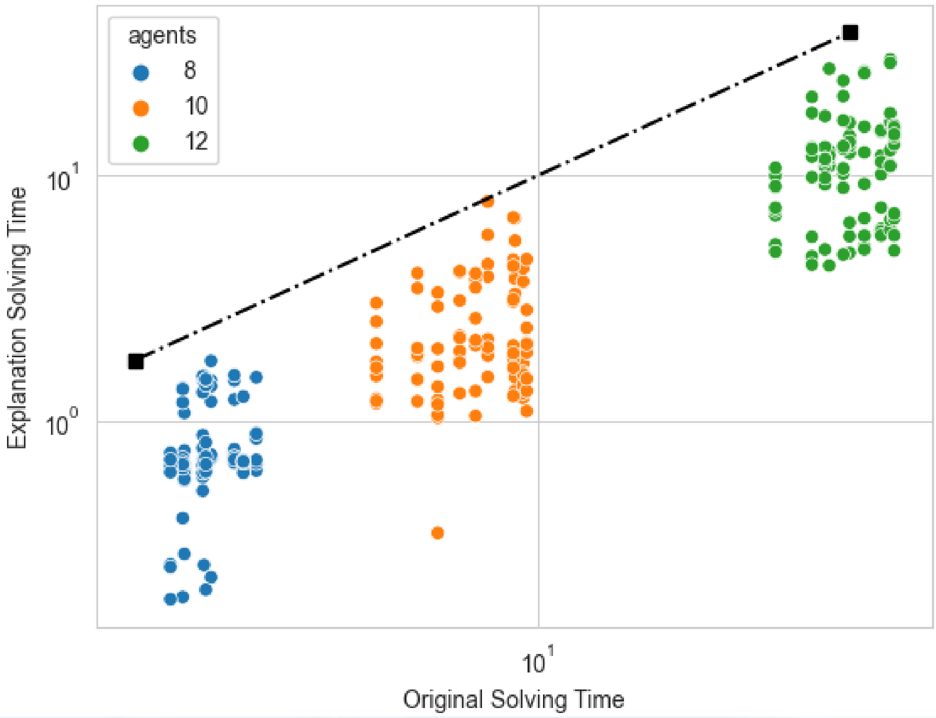


Knapsack

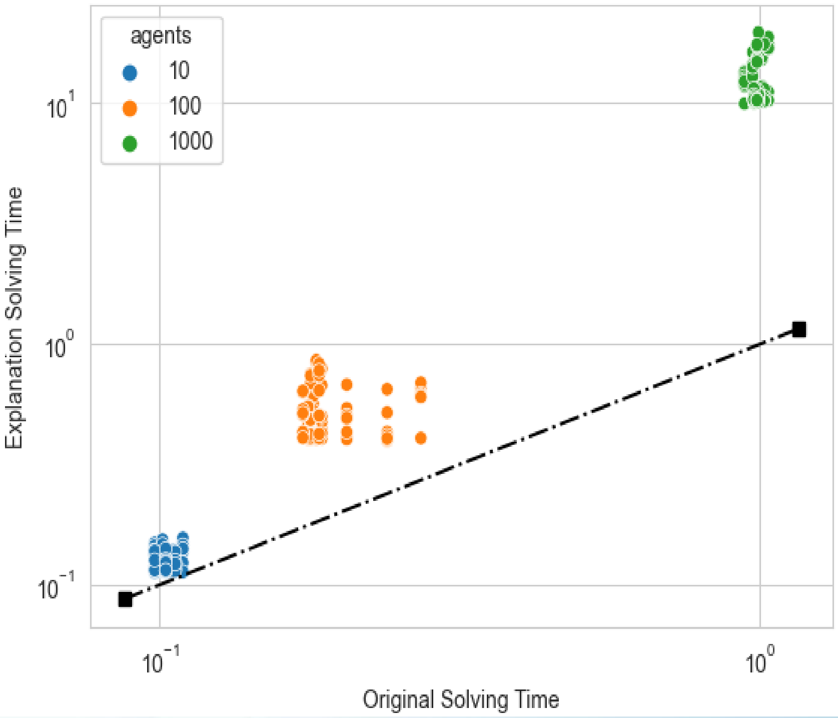


Solving Time: Original vs Explanation

Wedding Sitting



Knapsack



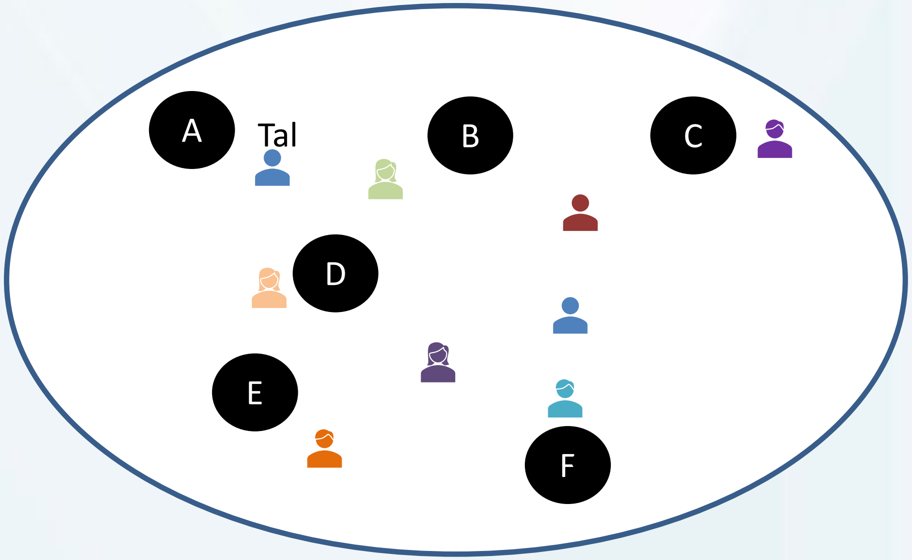
Results (208 subjects)

- Detailed and nondetailed explanations statistically significantly improve satisfaction and reduce complaints
- Explanation of blaming the algorithm does not make a difference.
- Detailed explanations are preferred over undetailed ones, but only in the VPR &Task Allocation domains were significantly “better” than non-detailed ones.

Explanation by a Mediator

- H1: An explanation will increase the willingness of human negotiators to accept a proposed agreement by an automated mediator agent.
- H2: An explanation will decrease the willingness of human negotiators to make a counteroffer to a proposed agreement by an automated mediator agent.
- H3: Humans prefer more detailed explanations.

Pick up location

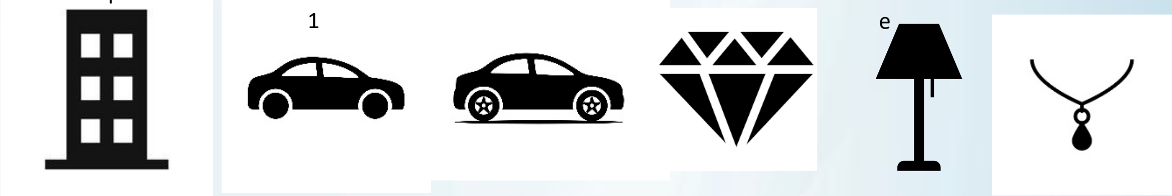


- Each agent has preference for each bus stop (A-F)

<u>Objective</u>: Reaching agreement on the location of the bus stop.

Inheritance division

- Each agent has preference for each item
- Each item has a different utility for each agent



Space: A Utility: U	Space: B Utility: V	Space: C Utility: W	Space: D Utility: X	Space: E Utility: Y	Space: F Utility: Z

<u>Objective</u>: Reaching agreement on how to divide the items between the agents.

Results (57 subjects)

- Detailed and nondetailed explanations statistically significantly increased acceptance and reduced the likelihood of a counter-proposal
- Most effective explanations varied among individuals and depended on the scenario.

Why do explanations help? -- they are cheap talk? (ongoing work)

- "Total friendship will decrease"
- Can we develop formal models that will yield strategies for agents that interact with humans and include explanations?

Yonatan Aumann (BIU)

Fairness matters

- **The utility function includes fairness consideration:**

1. What the agent believes is fair
2. What the agent believes others believe is fair
3. What the agent believes others’ act-upon fairness (norm)
4. What the agent thinks others will think about his behavior
5. The importance the agents assign to 1-4.

- **The utility is affected by the messages and actions**

Utility function with fairness

Utility function $U(v(z), f(\alpha, \beta, g, E))$ where:

- z: is the amount it keeps for itself.
- v: is the direct utility the aget obtains from share z.
- f: is the (dis)utility the agent obtains from an unfair distribution.
- g is an unfairness measure of the distribution.
- α: is a parameter reflecting the extent to which the agent dislikes unfair division.
- β: is a parameter reflecting the extent to which the proposer dislikes being judged by others as unfair.
- $E(\bar{\alpha}, \bar{\beta}, \bar{g})$ is the the belief of other agents α and g.

Example of a utility function

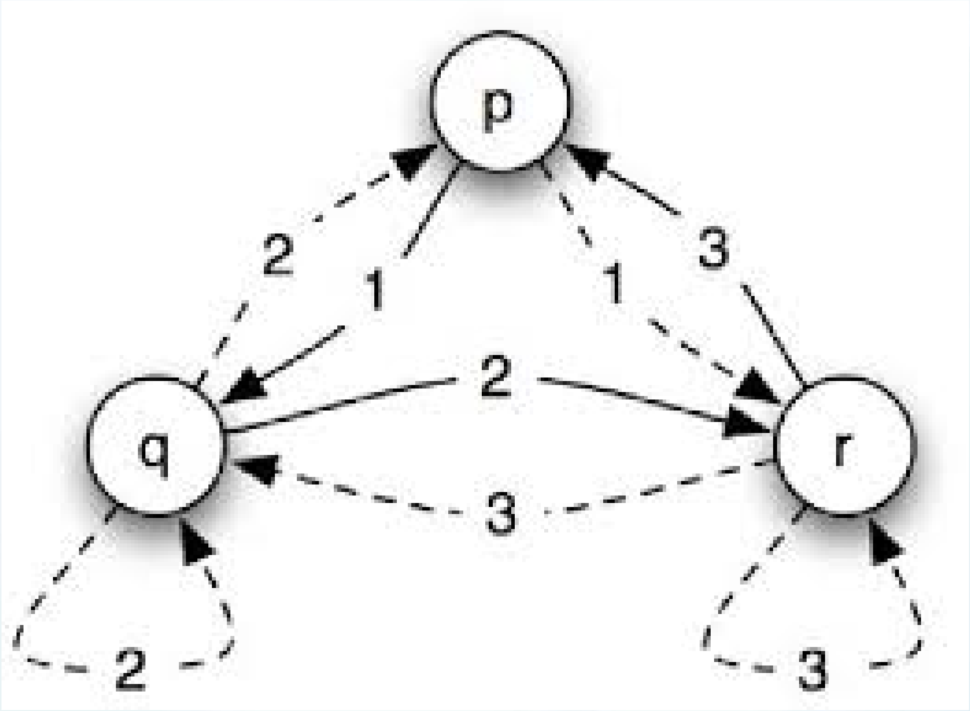
- f : $\alpha(\frac{g - E(b)}{\sigma(b)+1})$ g: agent adapted fairness

  E(b): agents' belief of the norm

- Role of explanations:
  - yield common belief of E(b)
  - change α

Possible world semantic

- $E(b)$: add a possible worlds model and use sequential Perfect Bayesian games.



Consideration of Cognitive aspects

- Change α: ???
  - Consideration action set/attention set to explain bounded rationality
  - **Consideration of cognitive aspects**, e.g. fairness, social welfare.

    We propose to incorporate it into the possible worlds.

  Messages change the current world of the agent.

Conclusions

- Explanations can change people's attitudes toward multi-agent solutions:
  - Providing information
  - Changing focus over cognitive and social consideration and beliefs about other agents.
- Running human studies is important to evaluate proposed social choice solutions.
- Development of formal models that take people's attitudes toward social norms is challenging but sheds light on people's behavior.