Model-Based Explanation for Automated Decision Making

David Garlan

Carnegie Mellon University

SoS Lablet Quarterly Meeting, November 2021

institute for SOFTWARE RESEARCH

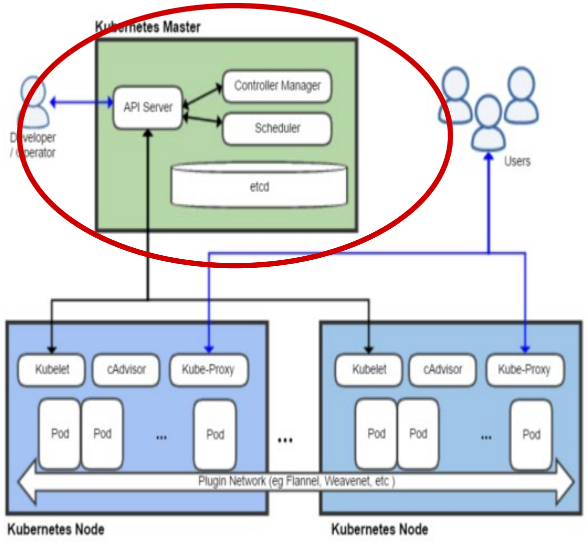
Carnegie Mellon

Collaborators

- Prof. Reid Simmons
  - Robotics, AI, planning, human-in-the-loop systems
- Rebekka Wohlrab
  - Requirements elicitation/negotiation, understanding tradeoffs
- Bradley Schmerl
  - Autonomous systems engineering
- Javier Camara
  - Probabilistic modeling, stochastic games, strategy synthesis
- Roykrong Sukkerd, Cody Kinneer, Ryan Wagner
  - PhD students

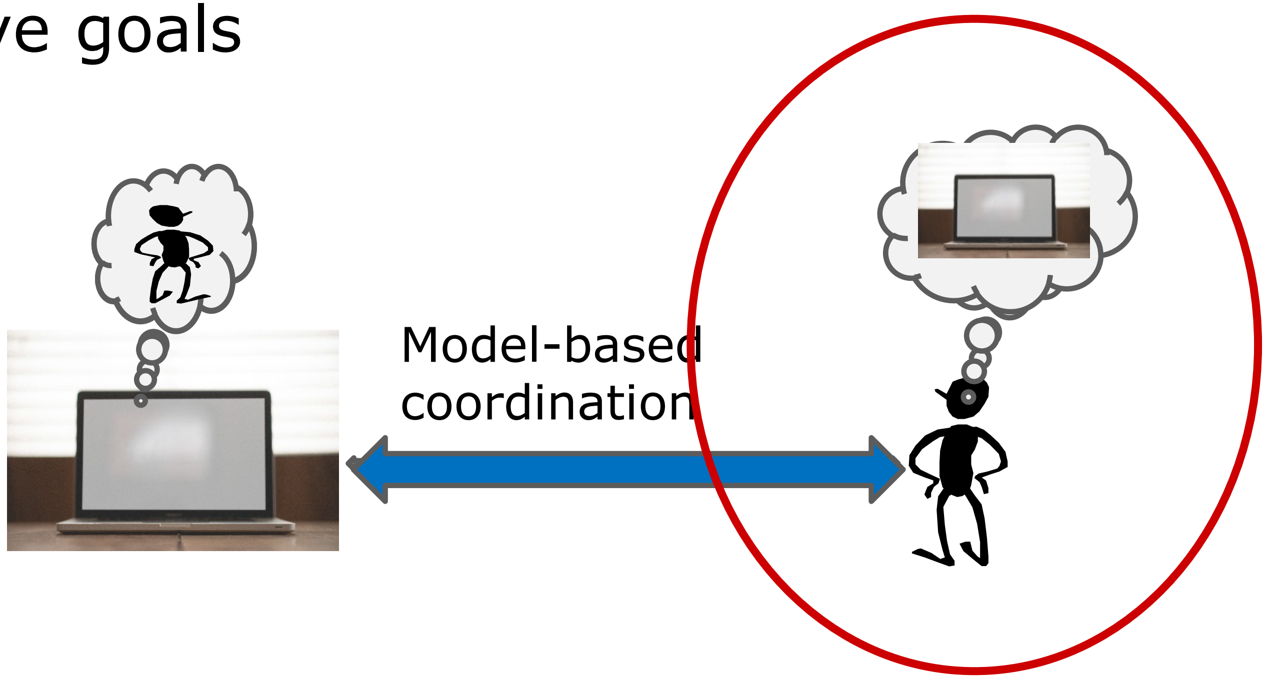
Problem Context

- Autonomy is increasingly important for modern systems
  - Widespread use in industry to manage faults, automate evolution, improve utility (e.g., Kubernetes)
  - Growing importance in managing security
- However, many systems require a combination of automated and human involvement to handle security attacks
- Problem: how to create effective coordination?
  - Decide which tasks to allocate to the system vs. human
  - Allow humans to have confidence in automated actions
  - Permit correction of erroneous or sub-optimal system actions
  - Improve automation by learning from what humans do
  - Understand what the system has done



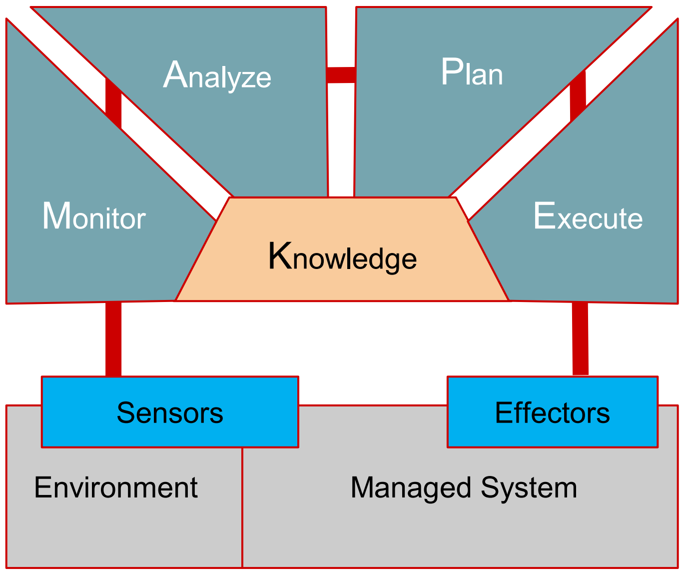
This Talk

- Context Recap
- Current Progress
  - Explanation of plans in presence of uncertainty and multi-objective goals
  - Contrastive explanations
  - User studies
  - Interactive explanation
  - Understanding the quality-attribute tradeoff space.

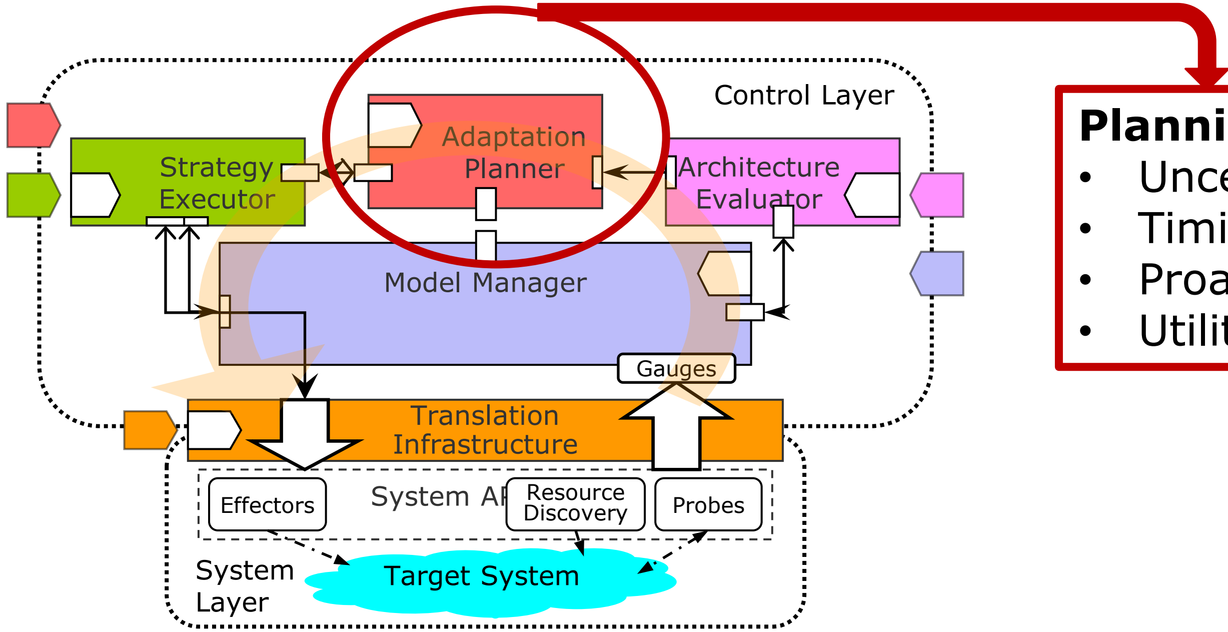


Context and Background

- In prior work we (and many others) have adopted a control systems view of system autonomy



MAPE-K

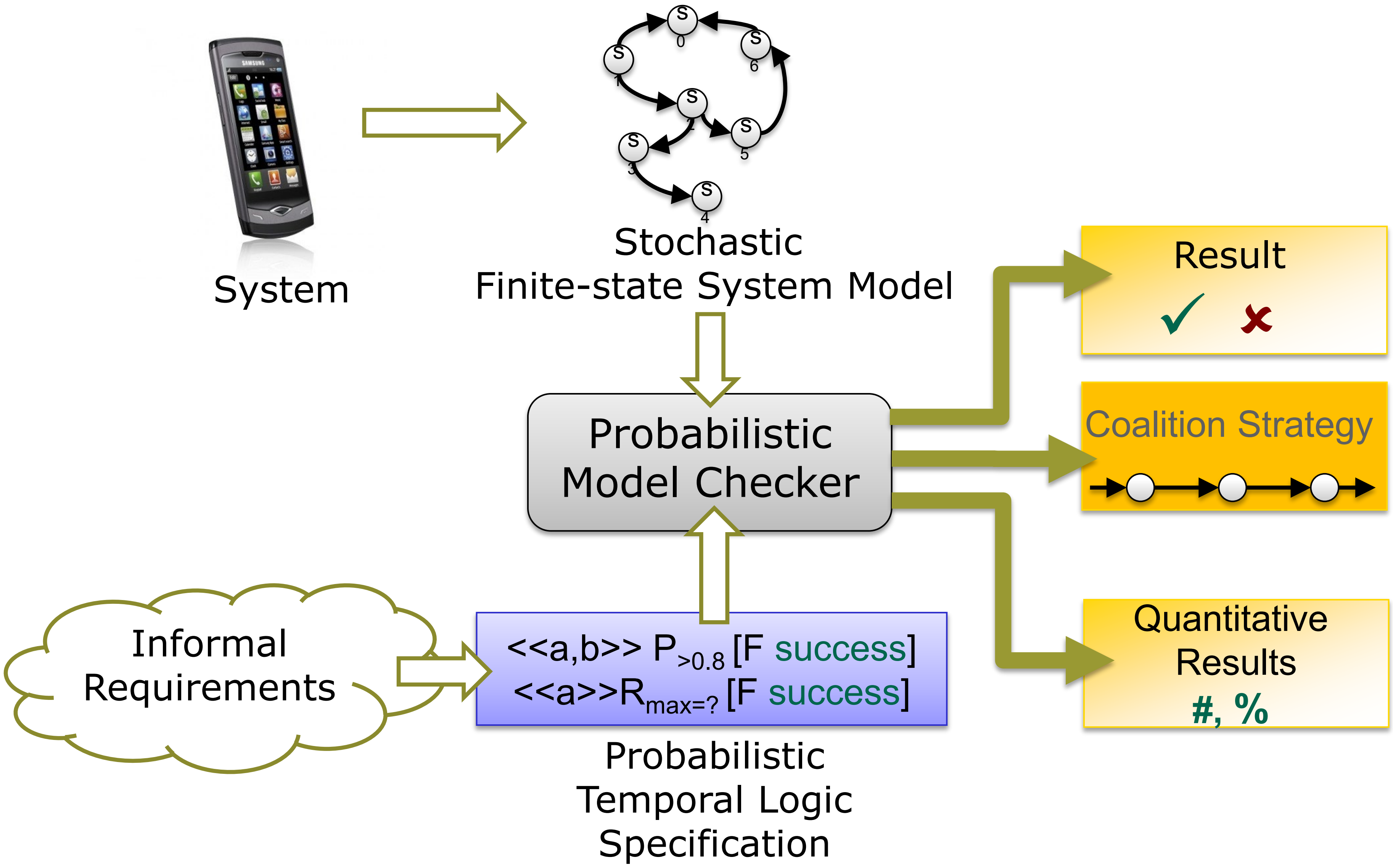


Rainbow Framework

Planning
- Uncertainty
- Timing
- Proactivity
- Utility

Probabilistic Models
Stochastic Games
ML-based

Formal Verification and Strategy Synthesis



Improving Transparency through Explanation

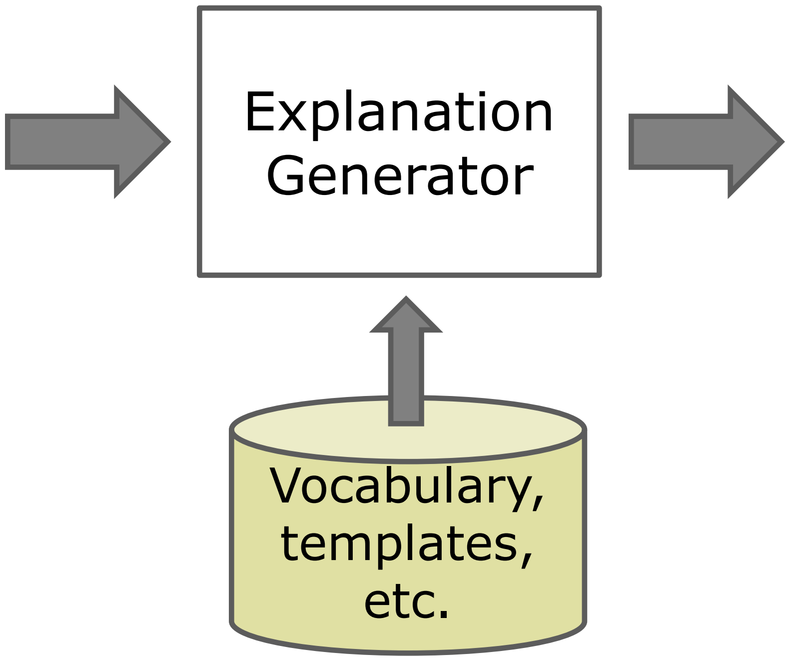
- Key idea: formal models for planning as the basis of human-understandable explanation.
- Elements of planning models for explanation:
  - Explicit goal for system adaptation
  - Explicit representation of quality dimensions and utility
  - Traceability from utility measures to quality dimensions and models that contribute to it.
  - Ability to explore alternative plans
  - Ability to interactively investigate alternatives
  - Understand the quality-attribute tradeoff space

Elements of an Explanation

- *"What am I trying to achieve?"*
  - Goal predicate, optimization objectives, constraints.
- *"What did I decide to do?"*
  - Narration of the chosen plan.
- *"What are the expected results and consequences of my decision?"*
  - Expected qualities and properties of the chosen plan (objective measures).
- *"What are some reasonable alternatives?"*
  - Select from a set of meaningful alternatives.
- *"Why did I reject other reasonable alternatives?"*
  - Value judgement and tradeoffs.
- *"What would be the consequences of changing my priorities?"*
  - Explore the tradeoff space.

Long-term Goal: A Generalized Tool for Explanation

- Plan
- Plan quality values
- Alternative plans
- Alternative plans' quality values



"I'm planning to **go through Corridor A** to **get to the target**. It would **take 2 minutes** and it would **have 0.05-probability of collision**. I could **reduce time to 1 minute**, but at the expense of **probability of collision (increase probability of collision to 0.2)**, by **going through Corridor B** instead. However, I decided not to do that because the **decrease in time** is not worth the **increase in probability of collision**."

Current Approach and Progress – Part 1

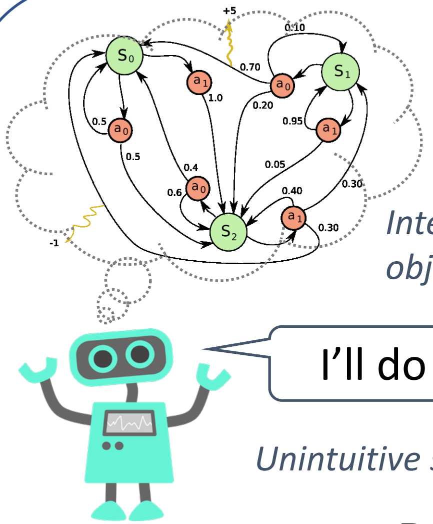
- Focus on task-oriented autonomy
  - Plans are generated to reach an explicit goal
  - Examples: robot navigation, responding to a particular kind of security attack
- Exploit explicit representation of utility to offer human-understandable explanations of a plan generated to solve a particular task
  - Utility: safety, timing, resource conservation, …
  - Tradeoffs: can reach a destination faster, but with a larger likelihood of crashing
- Present "reasonable" alternatives and why they were rejected
  - Contrastive or counter-factual explanations
  - Evaluate effectiveness of contrastive explanations through
  - User studies
- Allow a user to iteratively elaborate alternative possibilities and get explanations (work in progress)

Current Approach and Progress – Part 2

- Explore the quality attribute tradeoff space and use ML-based data reduction techniques to understand correlations, thresholds, key decisions, etc.
- Rebekka Wohlrab will present this part.

Improving Transparency and Intelligibility of Multi-Objective Planning

Explainable Multi-Objective Planning



Long-term consequences of actions…

Why this plan?

Interactions among objectives…

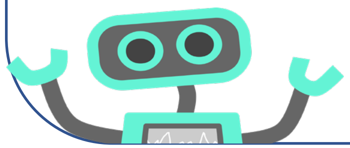
I'll do π.

Unintuitive solution…

<u>Problem</u>

Explainable Planning:

- Tradeoff-focused contrastive explanation
- Interactive explanation



<u>Approaches</u>

- Challenging for users to understand agent's rationale for its behavior
- May undermine user's trust, ability to collaborate with or correct agent

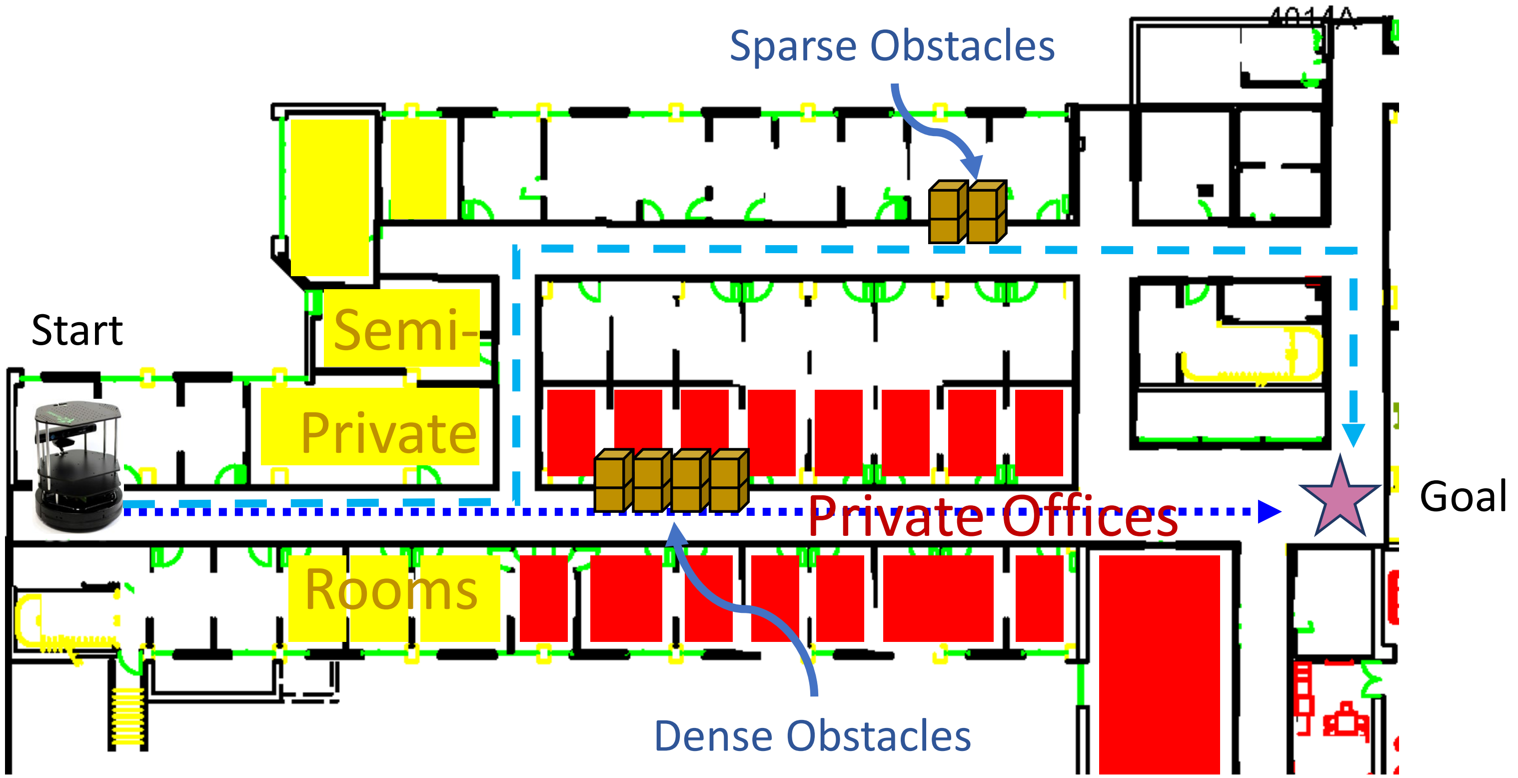
<u>Aims</u>

- Better understanding; higher confidence in assessing agent's decisions
- General framework

Tradeoff-Focused Contrastive Explanation

Consequence-oriented contrastive explanations for multi-objective Markov Decision Process (MDP) planning

Motivating Example



····→ is a more direct path, but
– – → is less intrusive and has lower chance of collision

Minimize:

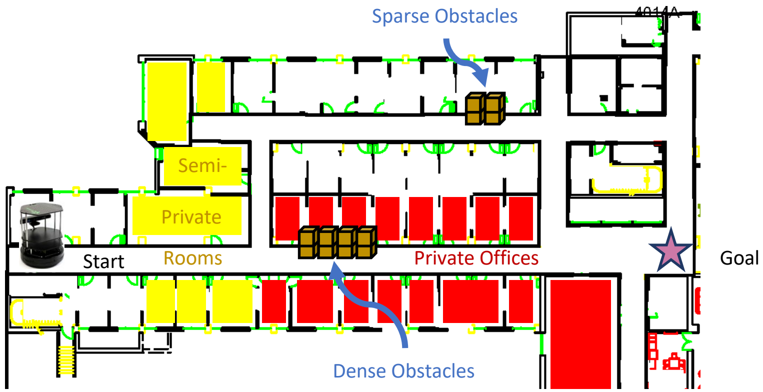
- Travel time
- E[Collisions]
- Intrusiveness

Quality Attributes (QAs)

I'm planning to follow – – → path. It is expected to take 5 minutes, have 0.2 expected collision, and be somewhat intrusive.

I could reduce the travel time to 4 minutes by following ····→ path instead. However, this would increase the expected collision to 0.4 and be very intrusive. I decided not to do that because the reduced time is not worth the increased expected collision and intrusiveness.

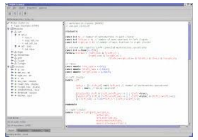
Explainable Planning Approach: Overview



Robot's Task

formulate as

$$\mathbf{P} = \begin{array}{c} \\ s1 \\ s2 \\ s3 \end{array} \begin{array}{c} \begin{array}{ccc} s1 & s2 & s3 \end{array} \\ \left[\begin{array}{ccc} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.3 & 0.6 \\ 0.8 & 0.2 & 0 \end{array}\right] \end{array}$$

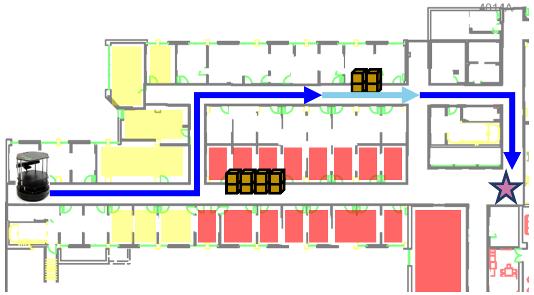


MDP Problem Representation

input

Planner

output



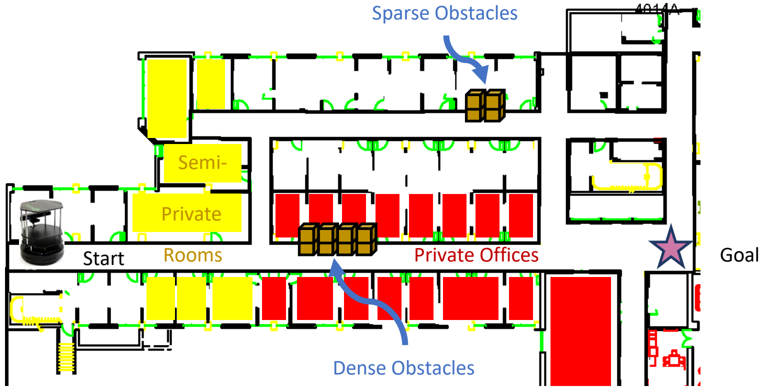
Navigation Plan

I'm planning to [follow this plan]. It is expected to [have these QA values].

I could [improve these QAs by these amounts], by [carrying out this alternative plan] instead. However, this would [worsen these other QAs by these amounts]. I decided not to do that because [the improvement in these QAs] is not worth [the deterioration in these other QAs].

Explanation

Explainable Planning Approach: Overview



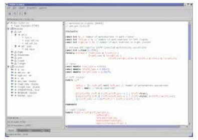
Robot's Task

formulate as

Explainable Planning Problem

translate to

$$\mathbf{P} = \begin{array}{c} \\ s1 \\ s2 \\ s3 \end{array} \begin{array}{c} \begin{array}{ccc} s1 & s2 & s3 \end{array} \\ \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.3 & 0.6 \\ 0.8 & 0.2 & 0 \end{bmatrix} \end{array}$$

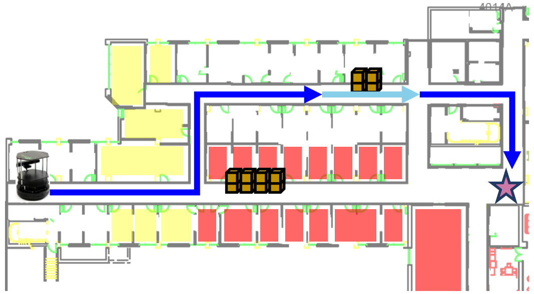


MDP Problem Representation

input

Planner

output



Navigation Plan

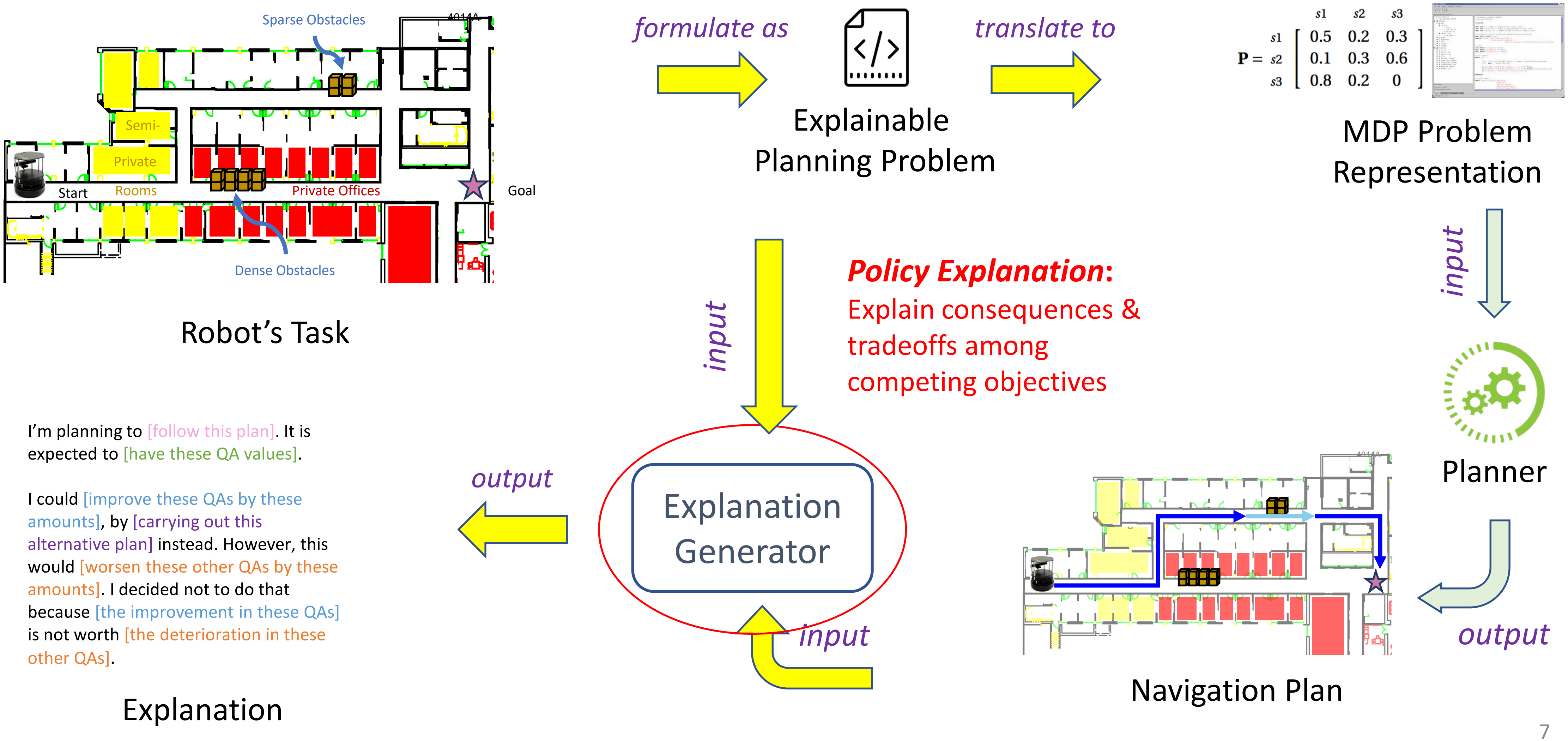
***Explainable Representation*:** Preserve the semantics of QAs

I'm planning to [follow this plan]. It is expected to [have these QA values].

I could [improve these QAs by these amounts], by [carrying out this alternative plan] instead. However, this would [worsen these other QAs by these amounts]. I decided not to do that because [the improvement in these QAs] is not worth [the deterioration in these other QAs].

Explanation

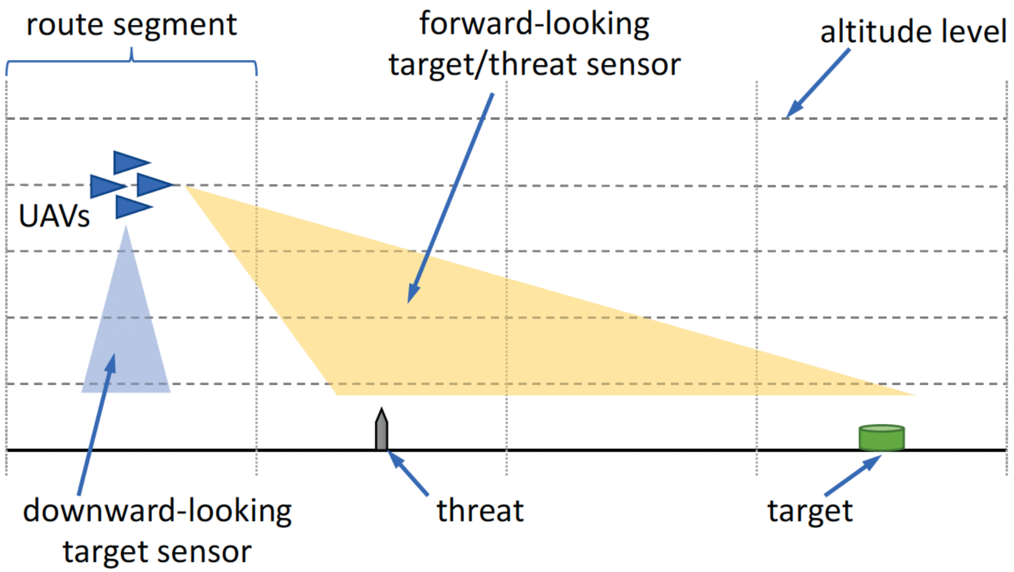
Explainable Planning Approach: Overview



General Applicability

Team of UAVs performing a reconnaissance mission in a hostile environment:

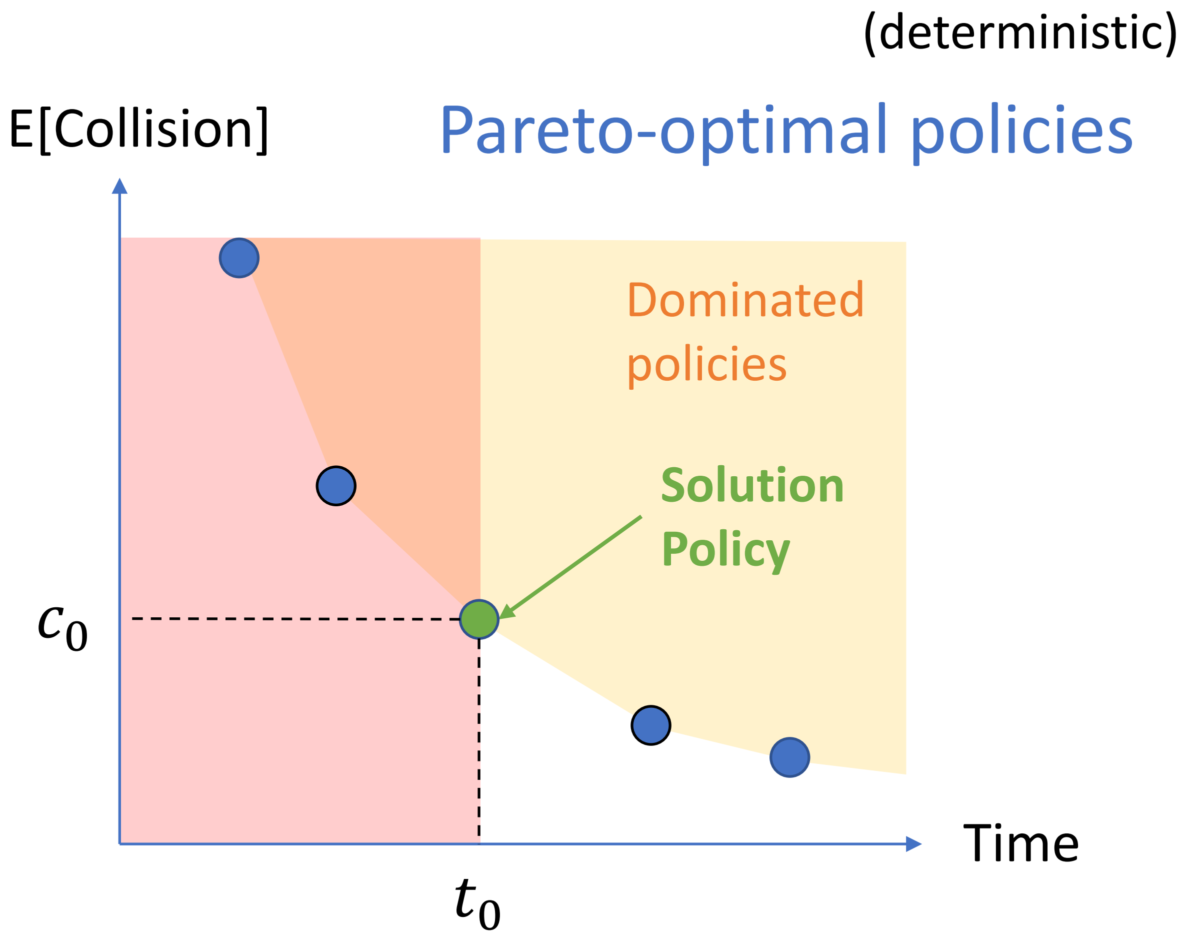
- Detect targets on the ground
- Avoid being shot down by threats



Outpatient clinic scheduling:

- Patient-related concerns: lead times
- Physician/system-related concerns: revenue, overtime and idle time

Consequences & Tradeoffs



Find tradeoff: **If & how** ***travel time*** is compromised for better *safety*

Find Pareto-optimal alternative policy whose Time < t_0

Find Alternatives: Constrained Planning

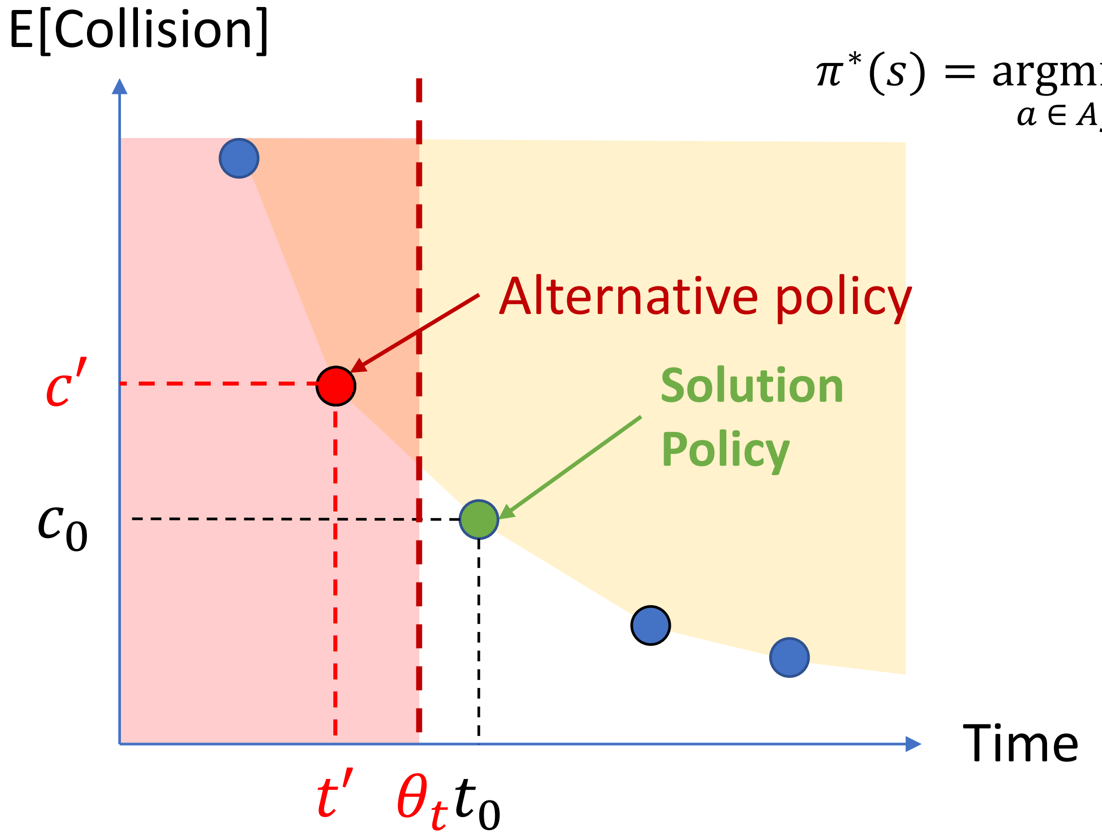
Demote "*Time*" objective

$$J(s) = \min_{a \in A_s} \left[\boldsymbol{C'(s,a)} + \sum_{s' \in S} Pr(s'|s,a)\, J(s') \right]$$

$$\pi^*(s) = \operatorname*{argmin}_{a \in A_s} \left[\boldsymbol{C'(s,a)} + \sum_{s' \in S} Pr(s'|s,a)\, J(s') \right]$$

subject to:

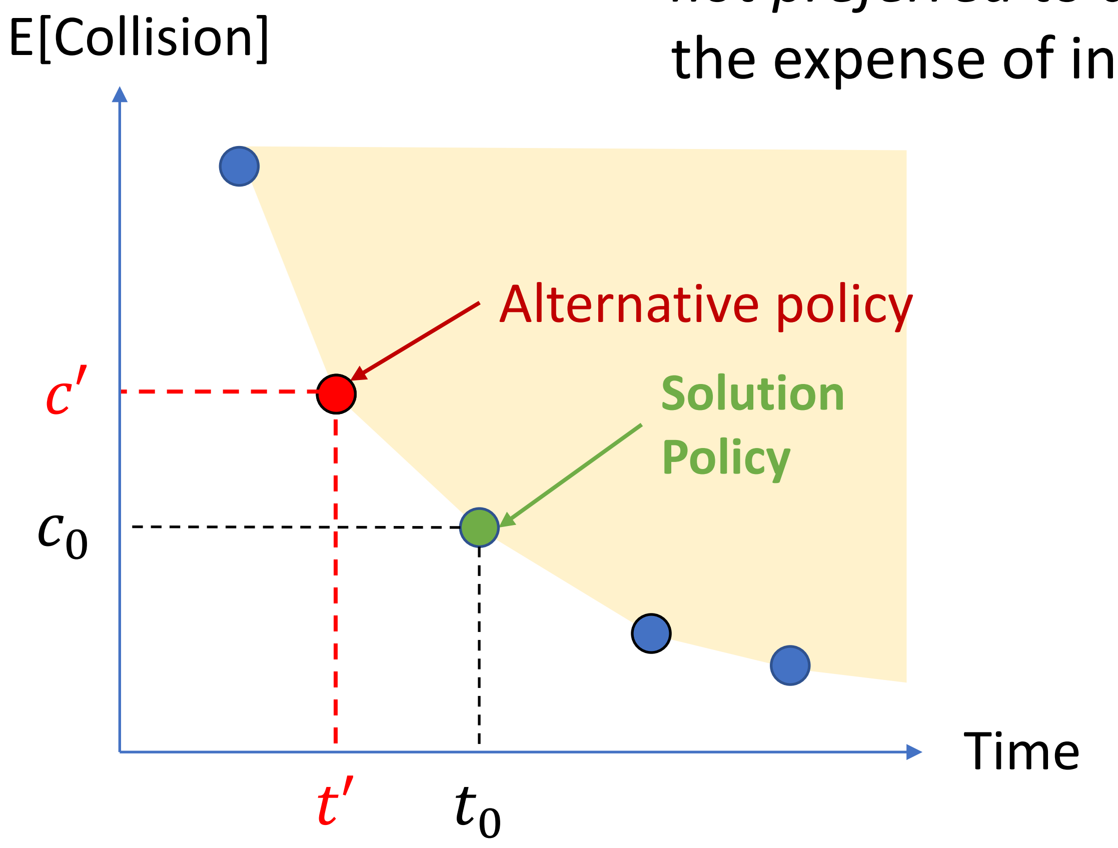
$$J^{\pi^*}_{time}(s_0) \leq \theta_t$$



Solve with MILP formulation that ensures ***deterministic solution***

Contrastive Explanation

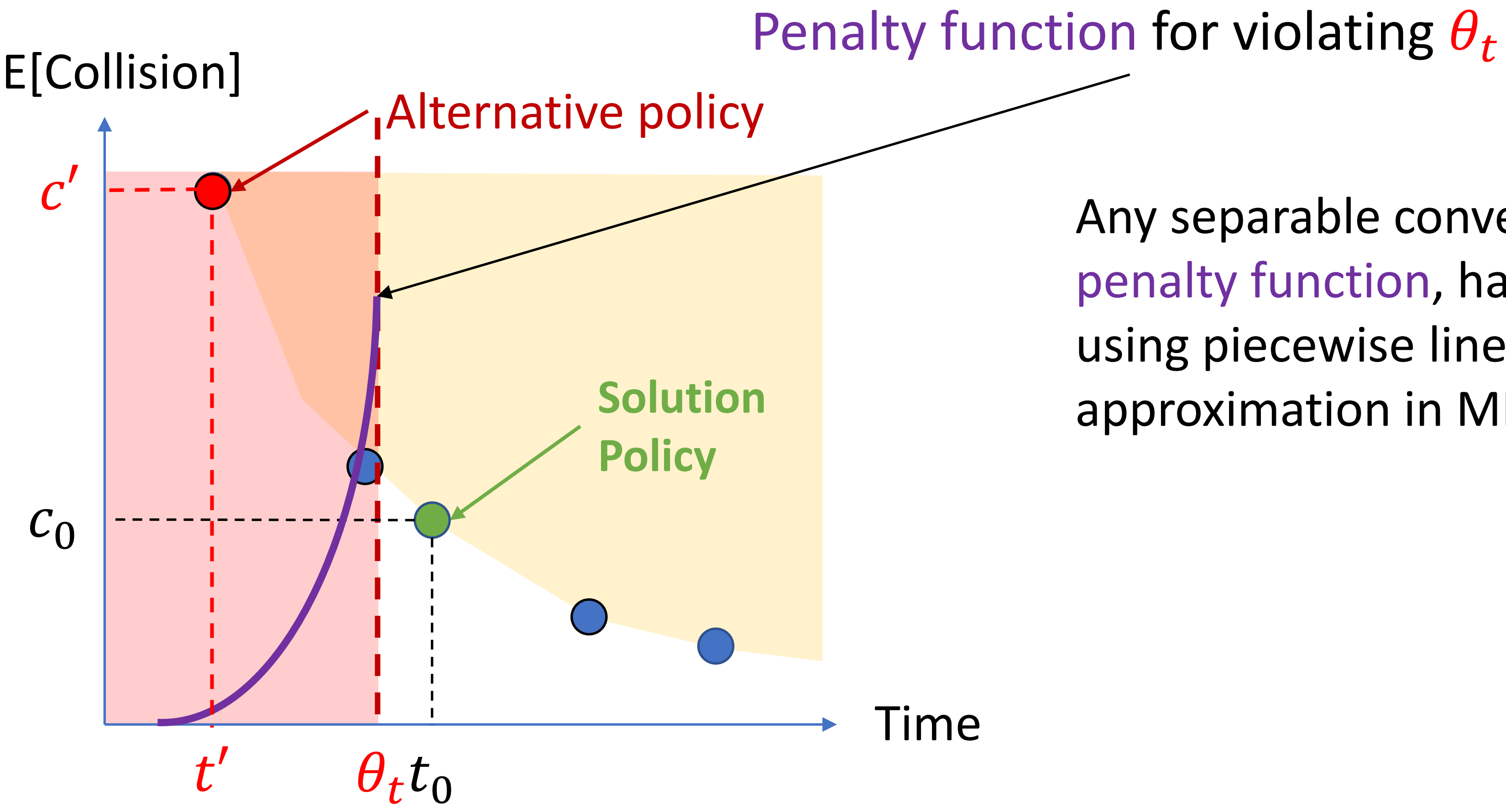
Explain the ***tradeoff rationale*** of decision: it is *not preferred* to decrease time from $\boldsymbol{t_0}$ to $\boldsymbol{t'}$ at the expense of increasing collisions from $\boldsymbol{c_0}$ to $\boldsymbol{c'}$



How ***travel time*** is compromised for better ***safety***

Find Alternatives: Soft Constraints

Alternatives that have too similar values don't show tradeoffs well



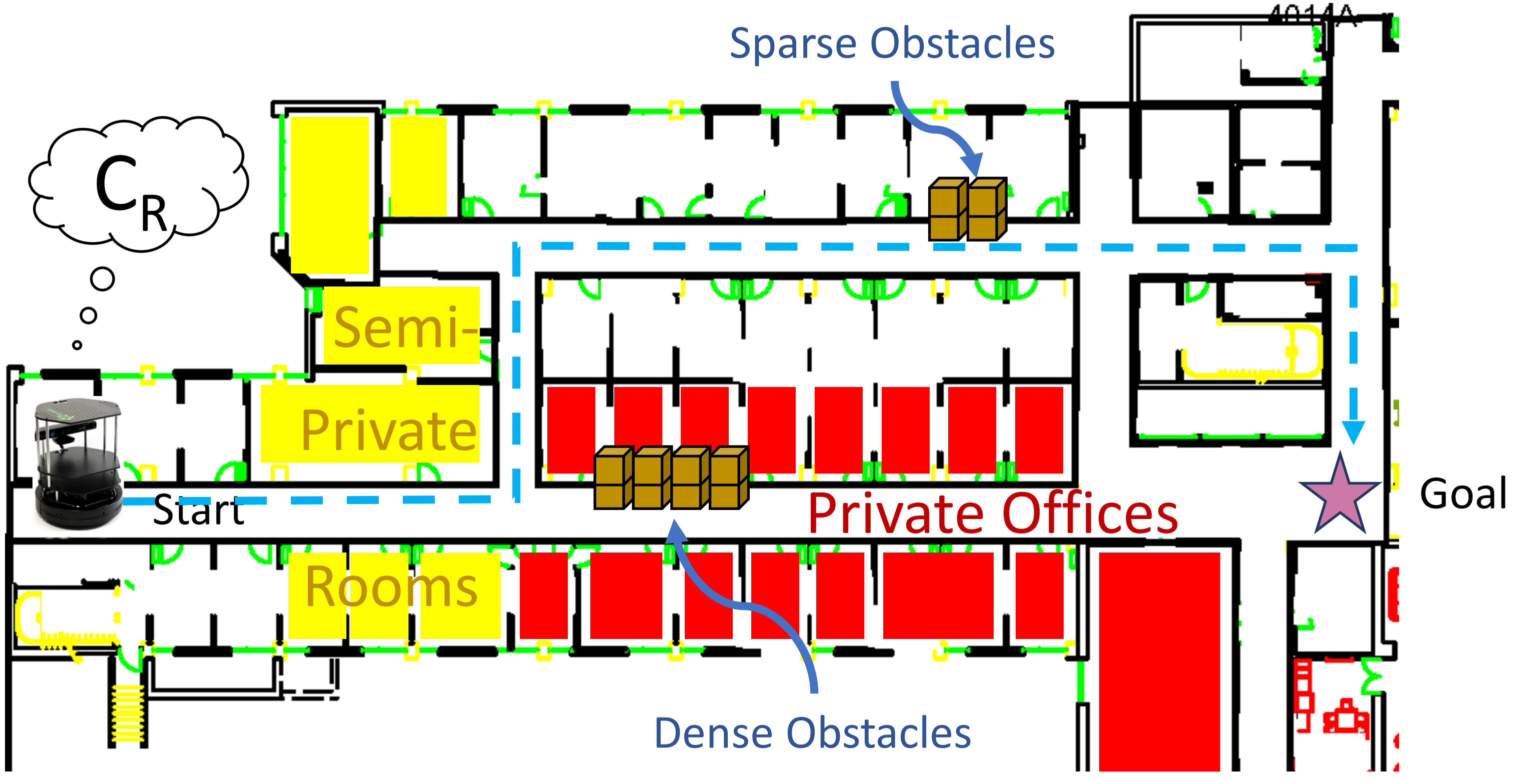
Any separable convex penalty function, handled using piecewise linear approximation in MILP

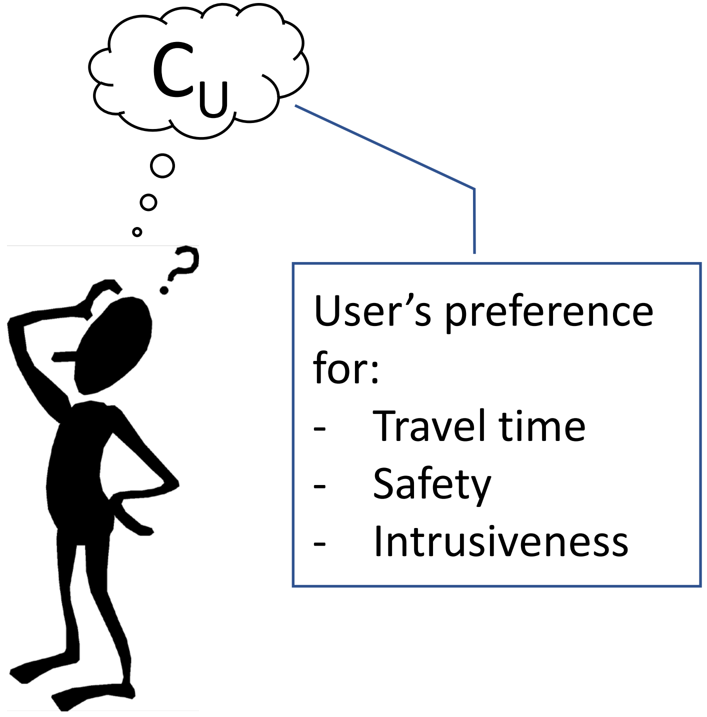
User Study Evaluation

Task-oriented human subjects study evaluation of explainable planning

User Study Scenario

Proxy for measuring understanding:



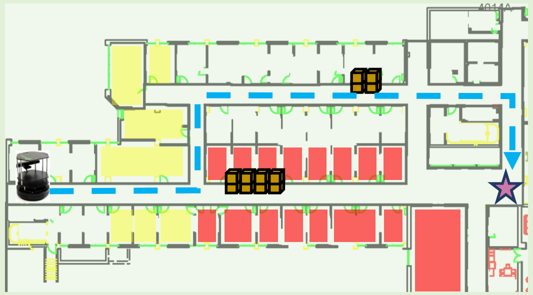


C_R may or may not be C_U

"Is the robot's plan the best option for me?"

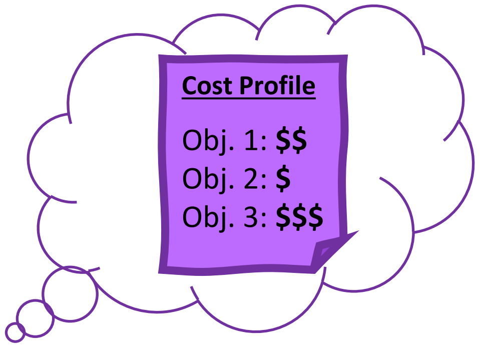
Control Group

Scenario



Robot's Navigation Plan

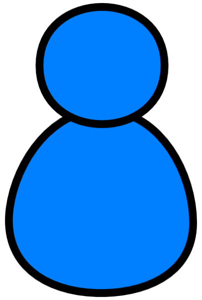
I'm planning to follow path.
It is expected to take 5 minutes, have 0.2 expected collision, and be somewhat intrusive.



Hidden from Participant

Cost Profile

Obj. 1: $
Obj. 2: $$$
Obj. 3: $$



Control Group

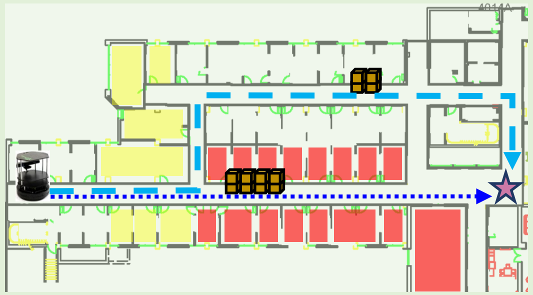
Participant

Is agent's plan the best option? [Yes/No]

How confident are you? [5-point Likert scale]

Treatment Group

Scenario



Robot's Navigation Plan

I'm planning to follow path. It is expected to take 5 minutes, have 0.2 expected collision, and be somewhat intrusive.

I could reduce the travel time to 4 minutes by following path instead. However, this would increase the expected collision to 0.4 and be very intrusive. I decided not to do that because the reduced time is not worth the increased expected collision and intrusiveness.

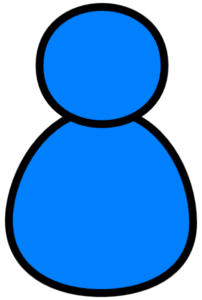
Cost Profile

Obj. 1: **$$**
Obj. 2: **$**
Obj. 3: **$$$**

Hidden from Participant

Cost Profile

Obj. 1: **$**
Obj. 2: **$$$**
Obj. 3: **$$**



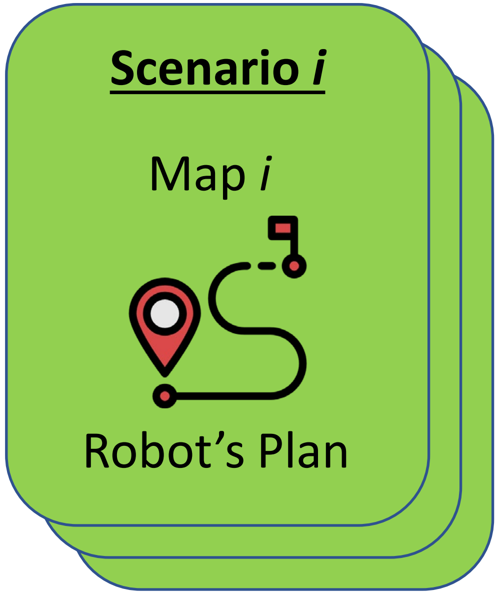
Treatment Group

Participant

Is agent's plan the best option? [Yes/No]

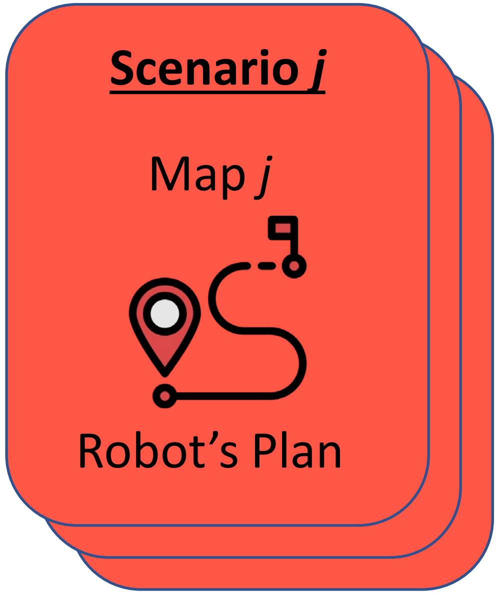
How confident are you? [5-point Likert scale]

Scenario Types



Preference-aligned
Scenarios

Agent's plan is the best option for user



Preference-misaligned
Scenarios

Agent's plan is *NOT* the best option for user

Hypotheses

H1: Participants who receive the explanations are more likely to correctly determine *whether the robot's plan is in line with their preference*.

H2: Participants who receive the explanations are more confident *in their determination*.

Results: Correctness

Mixed-Effect Logistic Regression: account for random effects from participants, scenarios

Correctness ~ Explanation, Scenario Type

H1 is supported

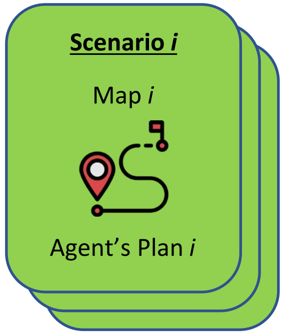
Treatment Group
(Given Explanation)
(49 participants)

is on average **3.8** times ***more*** likely to be *correct* with 95% CI: **[2.03, 7.12]**

Control Group
(50 participants)

Preference-misaligned
(24 scenarios)

is on average **0.36** times ***less*** likely to be *correct* with 95% CI: **[0.19, 0.70]**



Preference-aligned
(24 scenarios)

Results: Confidence

Mixed-Effect Linear Regression: account for random effects from participants, scenarios

Confidence ~ Explanation, Scenario Type

H2 is supported

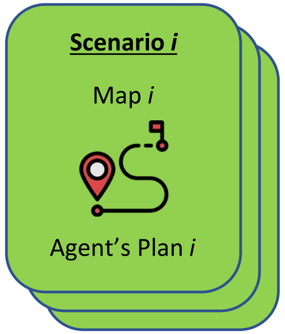
Treatment Group
(Given Explanation)
(49 participants)

is on average **0.42**
more *confident*
(Medium effect size: d=0.43)
with 95% CI: **[0.09,0.74]**

Control Group
(50 participants)

Preference-misaligned
(24 scenarios)

No statistically significant difference

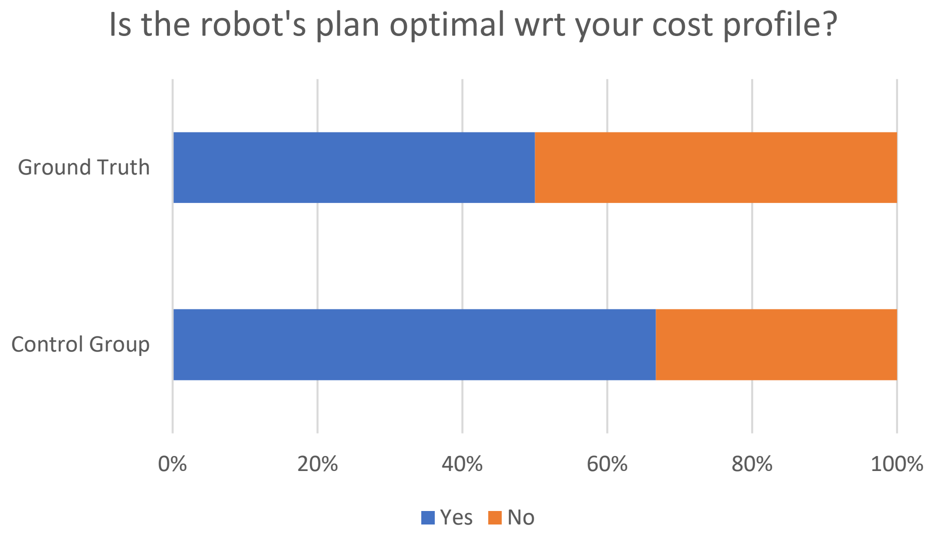


Preference-aligned
(24 scenarios)

Potential Overtrust When Unexplained

Participants have the tendency to agree with the robot's decisions, in absence of explanations.

Control Group

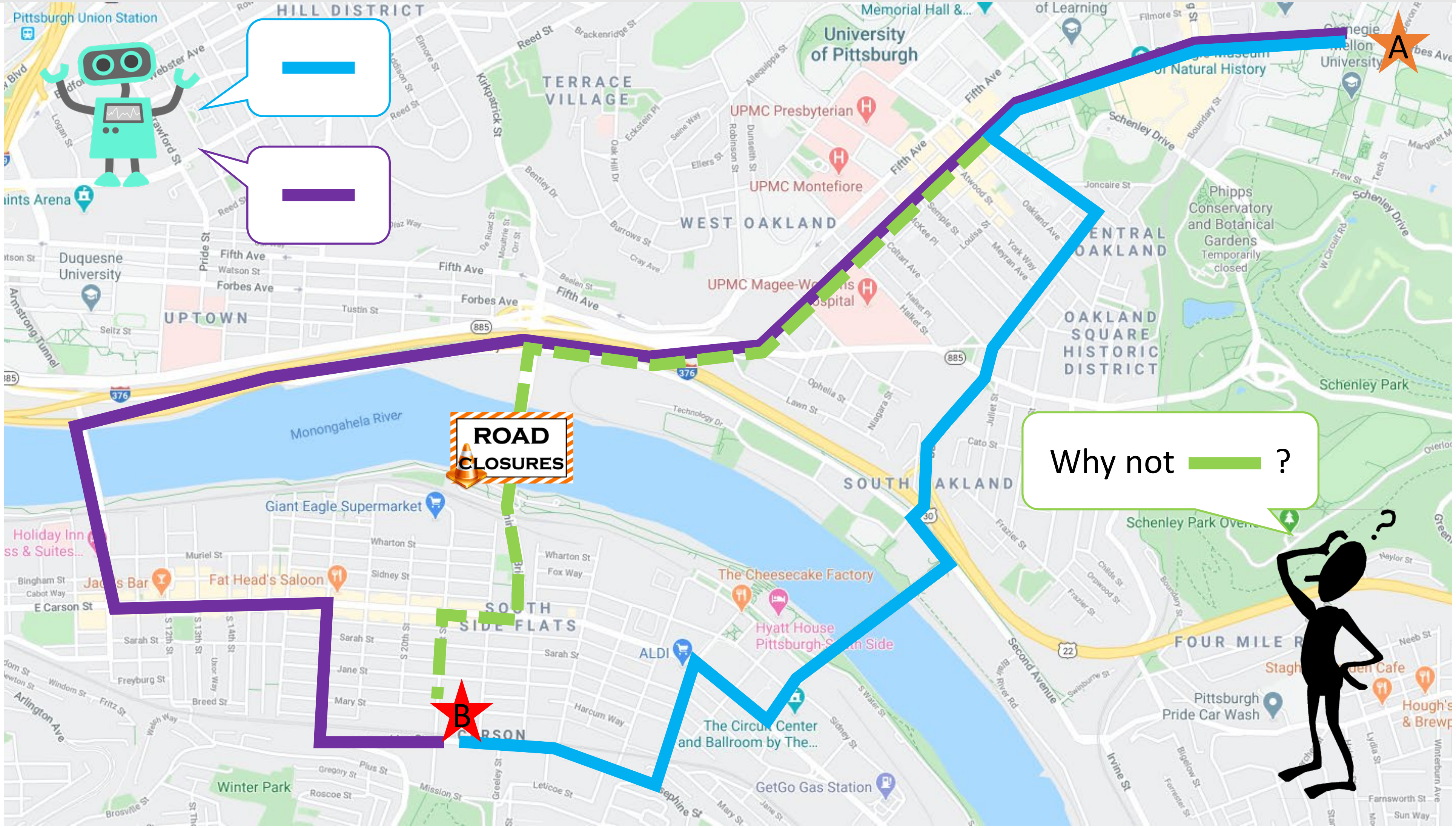


Bad news when the robot is misaligned with the user's preference.

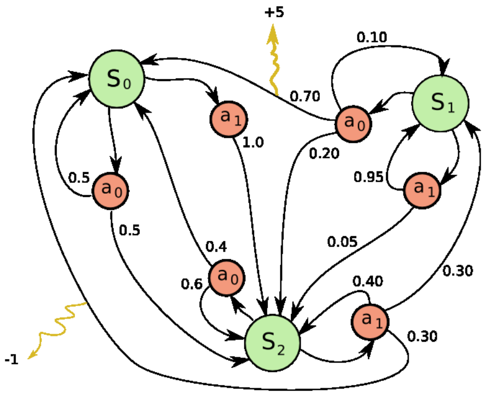
Interactive Explainable Planning

Interactive and iterative mechanisms for explainable planning

Address Unexpected Behavior

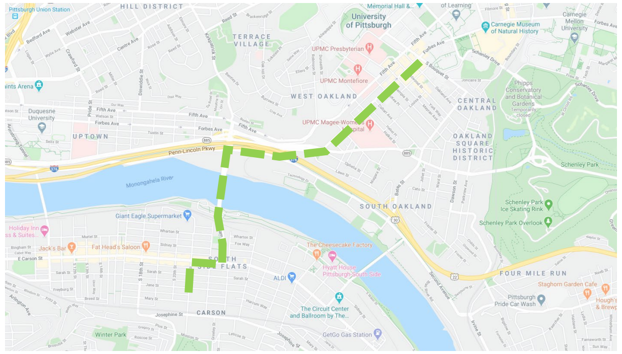


User Query as Planning Constraint



MDP

+

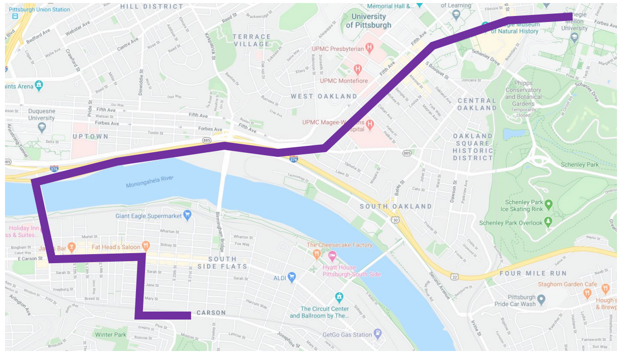


Constraint

Re-Planning

- Different types of constraints
- Different approaches to handle constraints

User-Guided Explanation



Alternative Solution

User Query as LTL Property

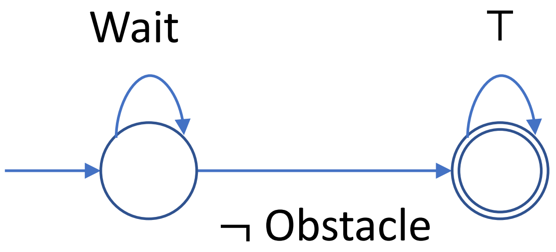
Linear Temporal Logic (LTL) formulas:

$$\varphi ::= true \mid a \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \mathrm{X}\,\varphi \mid \varphi_1\, U\, \varphi_2$$

Example: Robot should wait until somebody moves the obstacle out of its way.

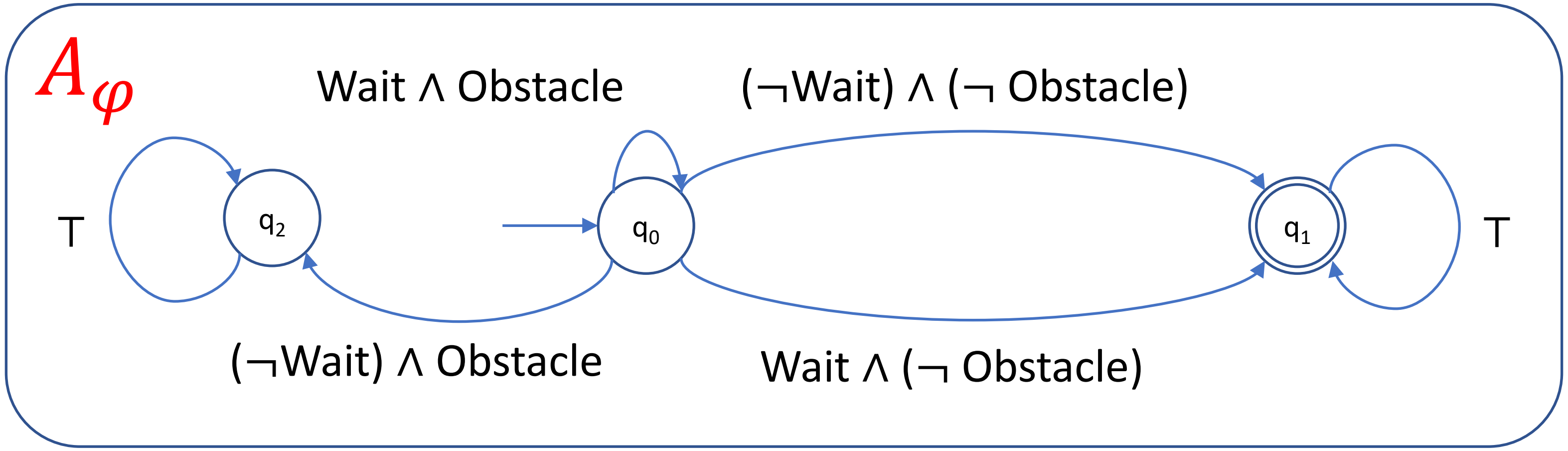
LTL specification: Wait ***U*** (¬ Obstacle)

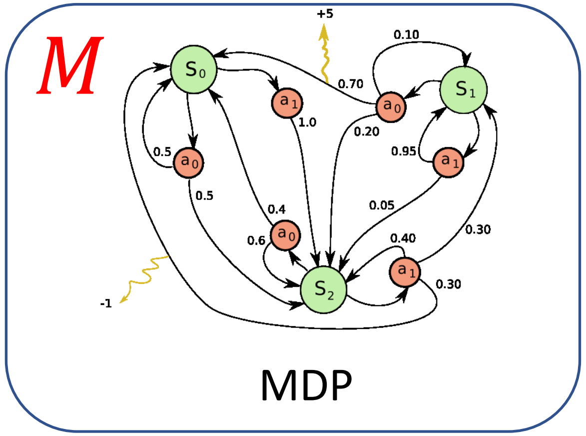
(Nondeterministic)
Büchi automaton:



Planning with LTL Constraint

Deterministic Rabin Automaton (DRA):





Product MDP: $M \otimes A_\varphi$

Planning: reaching acceptance condition

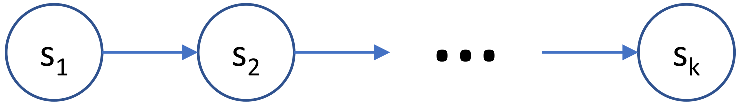
Constraint-Satisfying Policy

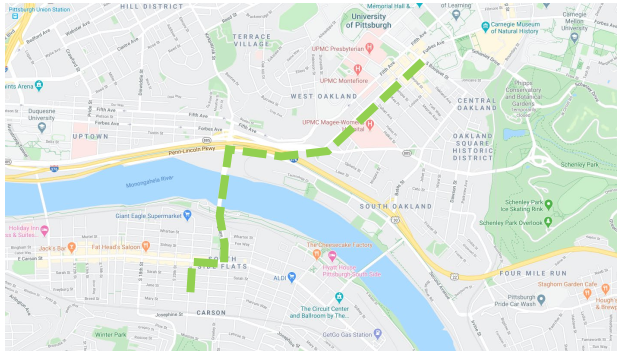
Explain consequences of query & tradeoffs

Handle Unsatisfiable Query

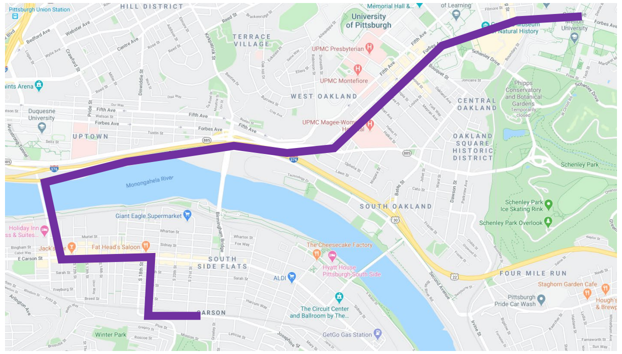
Maximum Realizability:

- *Simple case*: state trajectory constraint





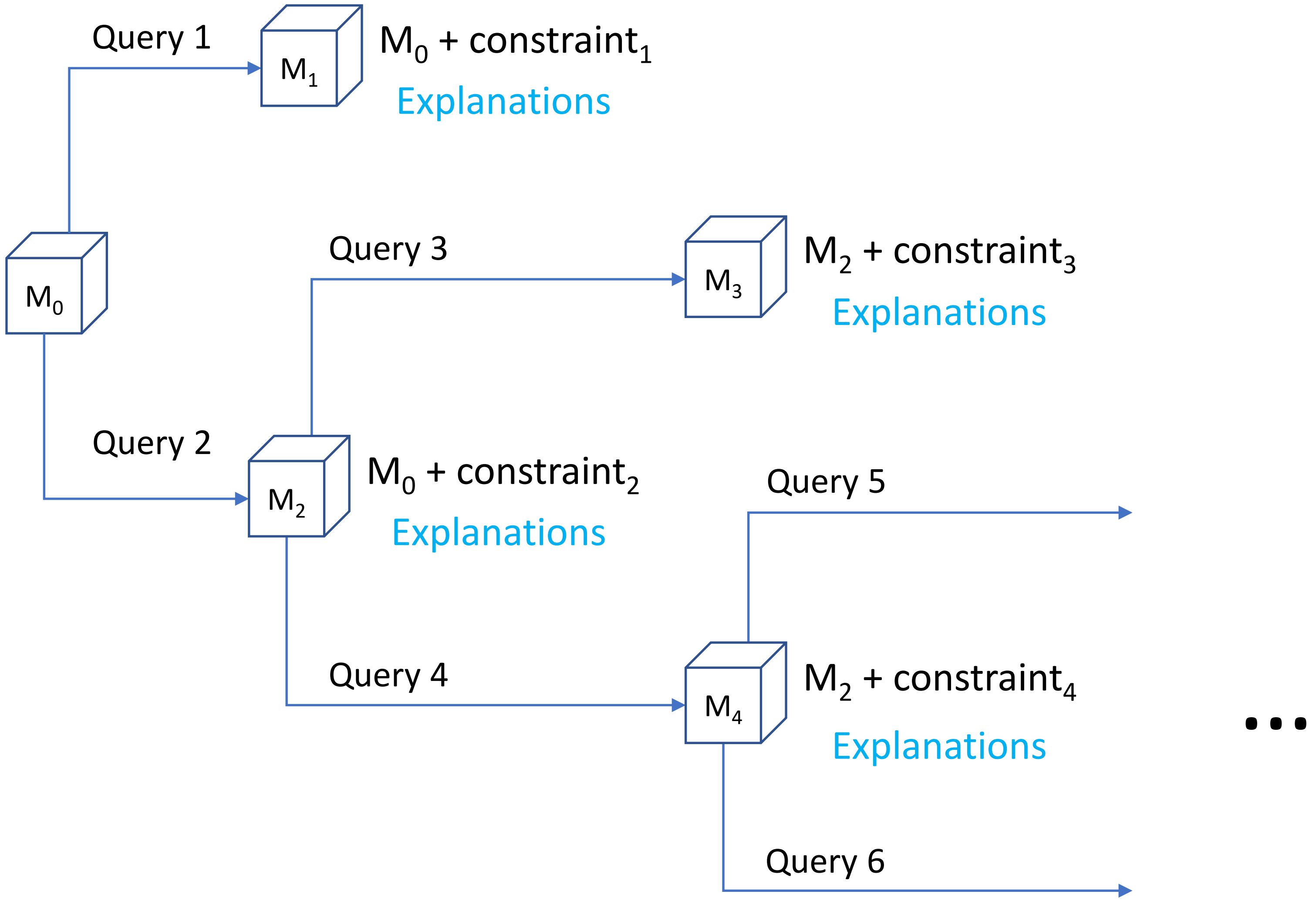
State Trajectory Constraint



Maximum Realization

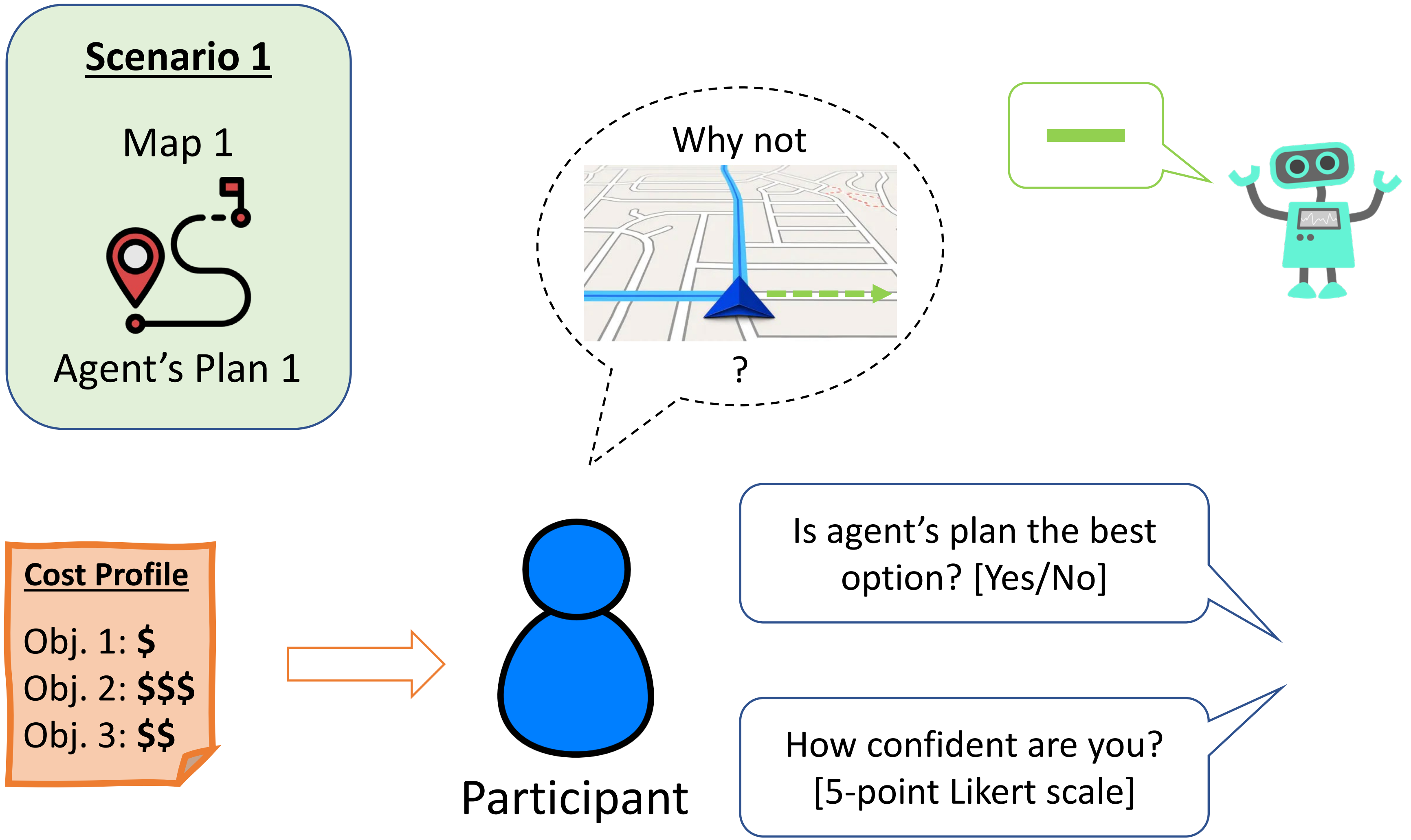
- *More general*: soft constraint $\square\,\varphi$ (future work)

Iterative Query & Explanation

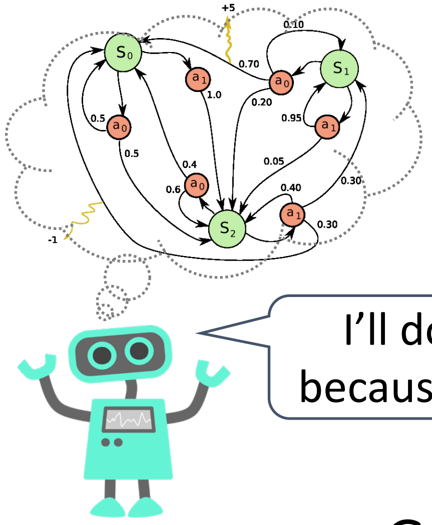


User can iteratively refine their queries to clarify their questions, to get refined explanations.

Evaluation of User-Guided Explanation



Summary

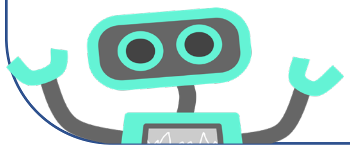


I'll do π because [...]

Goal

Transparency and intelligibility of multi-objective planning

Explainable Planning:
- Tradeoff-focused contrastive explanation
- Interactive explanation



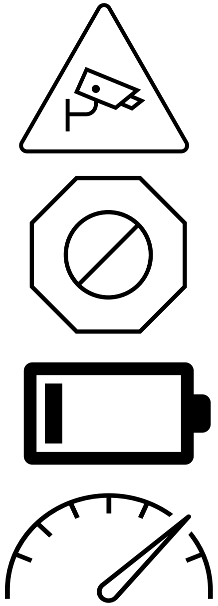
Approaches

Results
- Explanations improve understanding, confidence in assessing agent's decisions
- General framework

Rebekka Wohlrab

- Postdoc at the Institute for Software Research at Carnegie Mellon University, Pittsburgh
- Research interests: Requirements engineering, software architecture, self-adaptive systems, empirical software engineering
- PhD in Computer Science from Chalmers University of Technology, Gothenburg, Sweden
  - Thesis topic: Living Boundary Objects to Support Agile Inter-Team Coordination

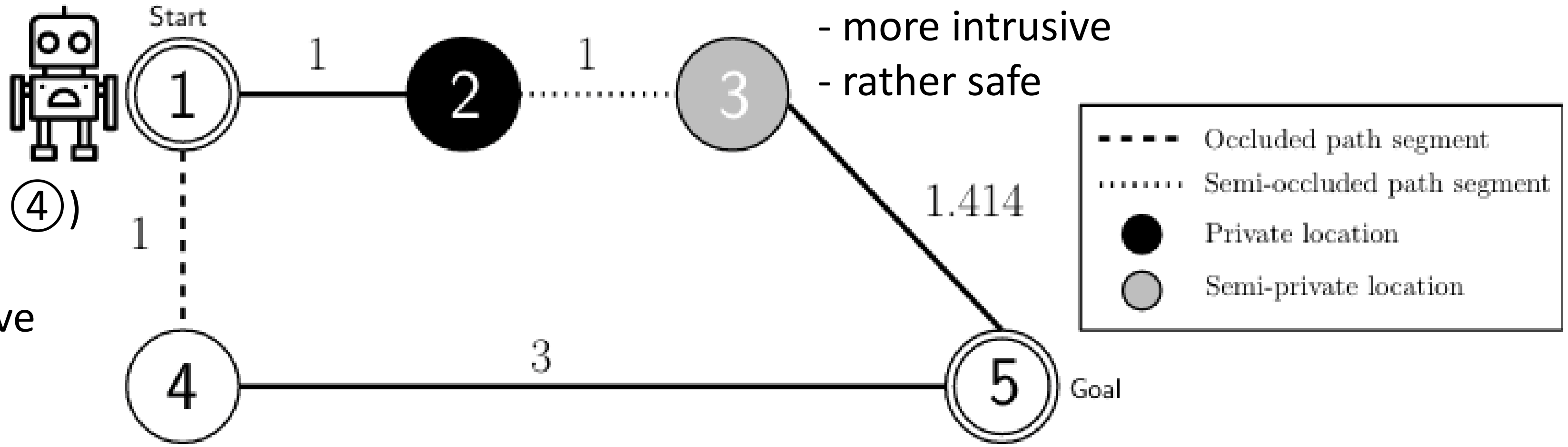
Quality Tradeoffs for Self-Adaptive Systems

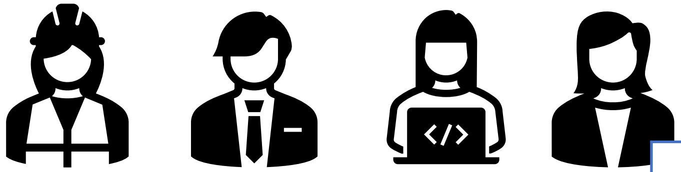


utility(plan) = 0.8·utility_travel_time(plan)+0.1·utility_safety(plan)
+0.1·utility_intrusiveness(plan)

Policy A (via ② and ③)
- shorter
- more intrusive
- rather safe

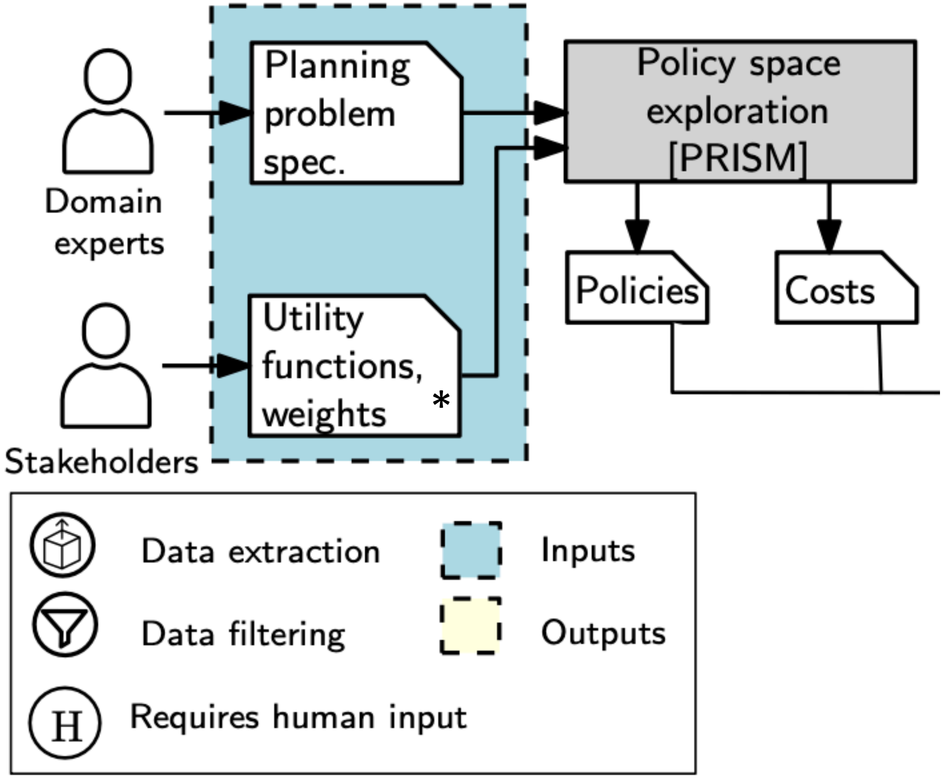
Policy B (via ④)
- longer
- not intrusive
- less safe



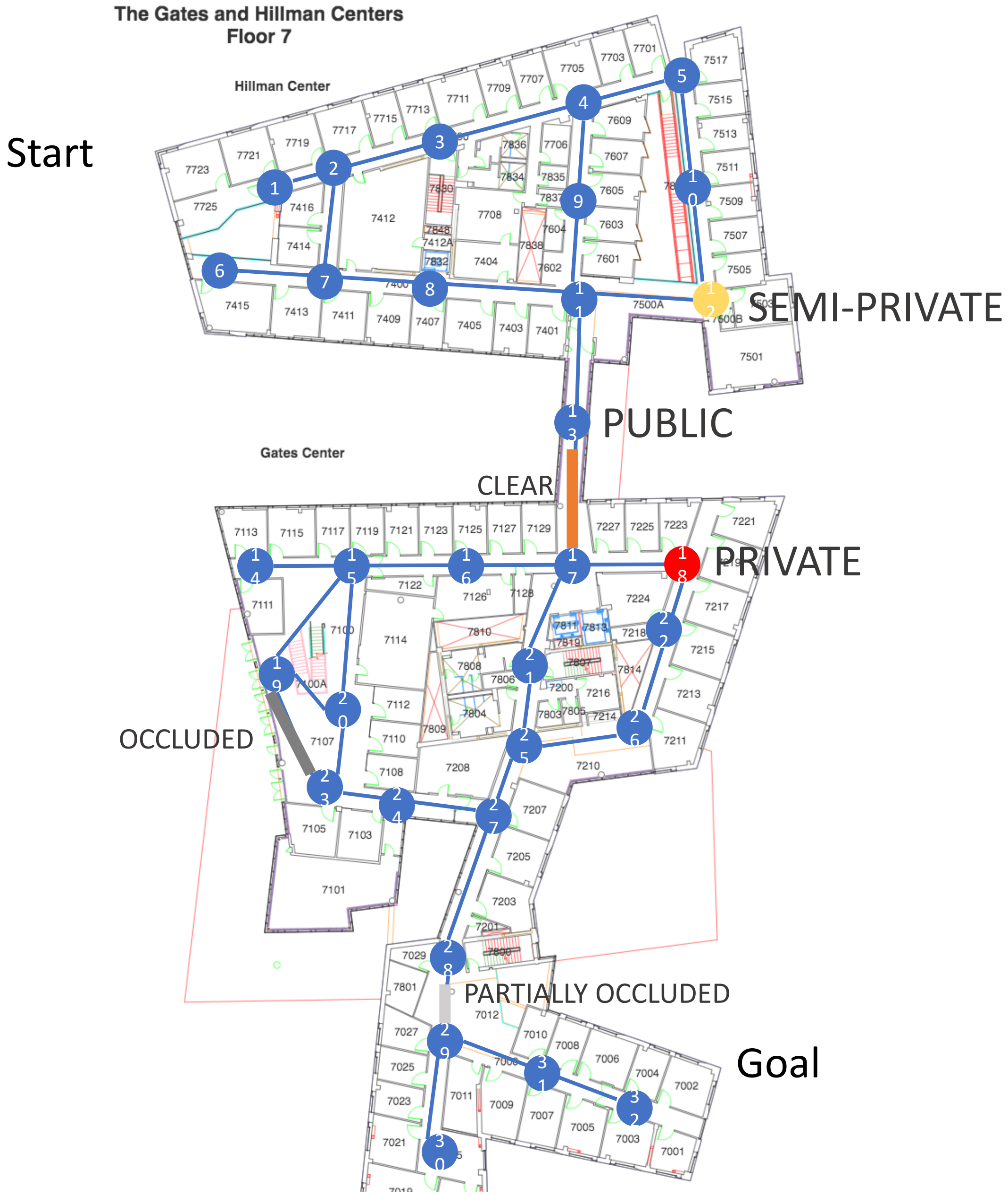


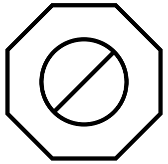
- Why are these policies being generated and not others?
- What are the underlying tradeoffs among quality attributes?
- Which are the key choices that drive the most important changes in adaptation behavior?
- What changes in the utility function would lead to different policies being generated?

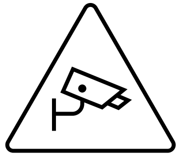
Overview of the quality tradeoff explanation approach

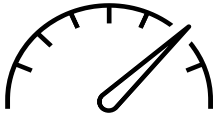


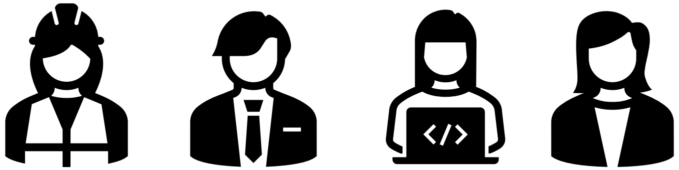
* R. Wohlrab and D. Garlan. A Negotiation Support System for Defining Utility Functions for Multi-Stakeholder Self-Adaptive Systems. Accepted to *Requirements Engineering*. 2021




Safety

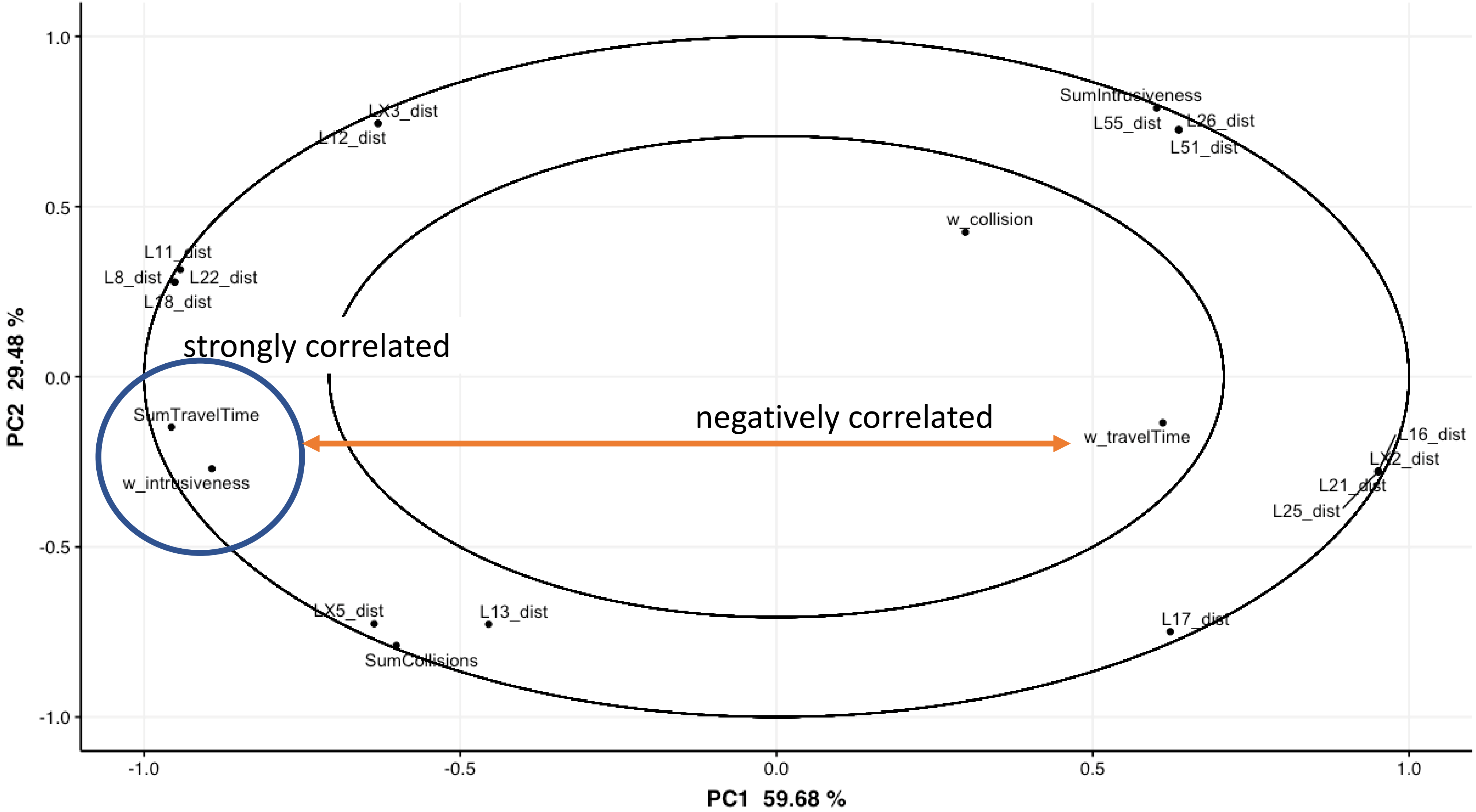

Intrusiveness


Travel time

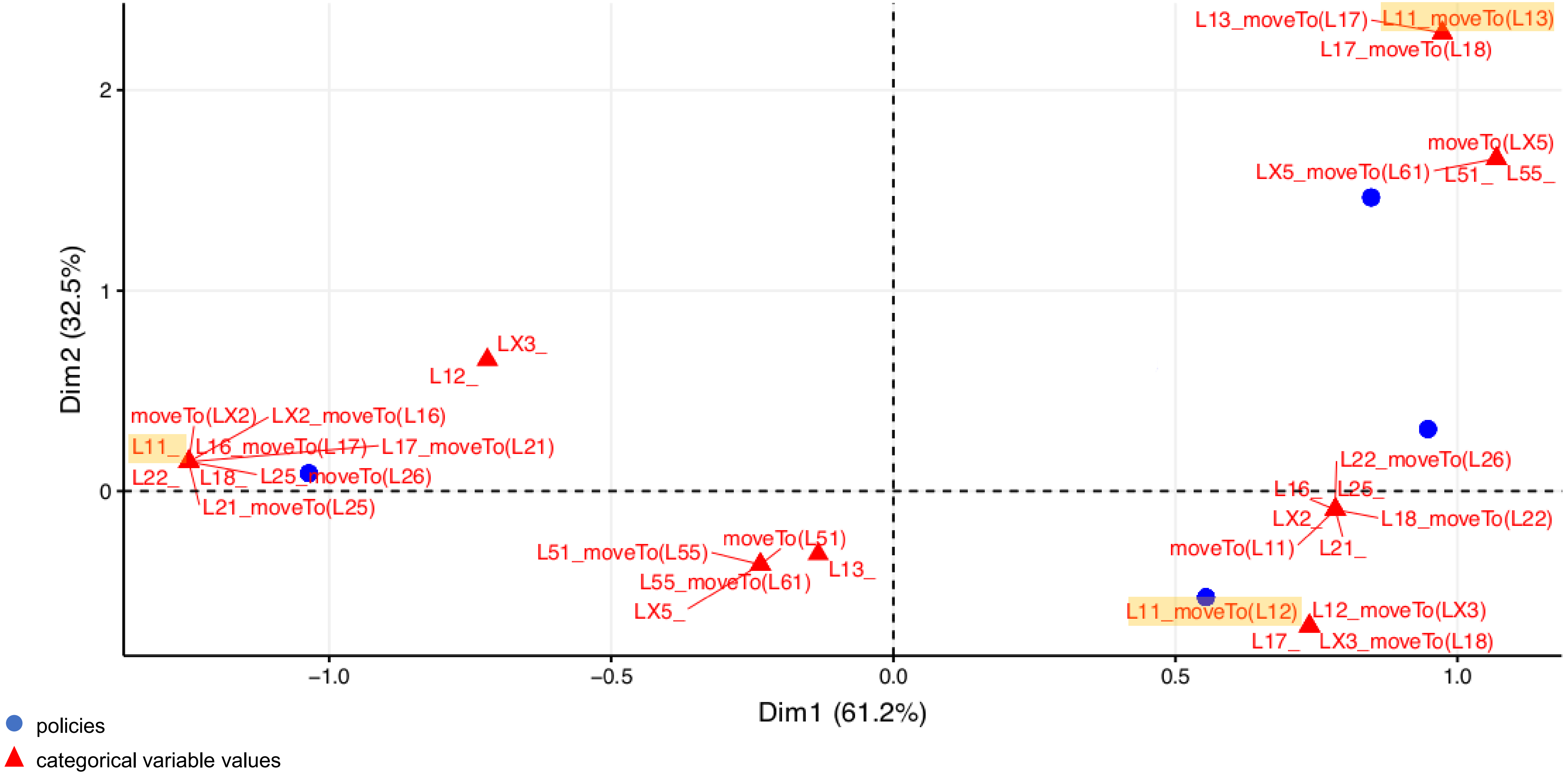


utility(plan) = 0.8·utility_travel_time(plan)+0.1·utility_safety(plan)
+0.1·utility_intrusiveness(plan)

Principal Component Analysis (PCA)

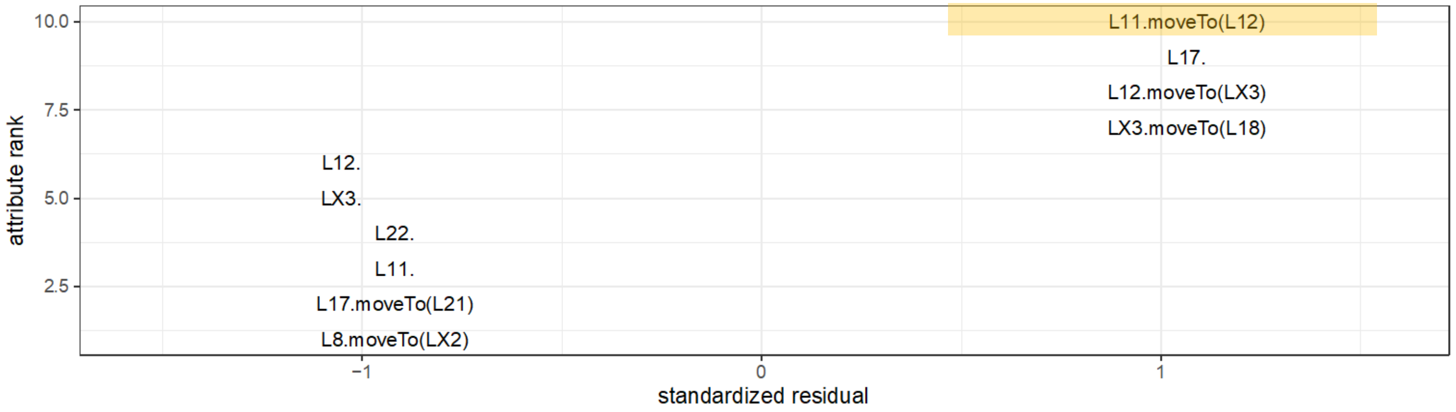


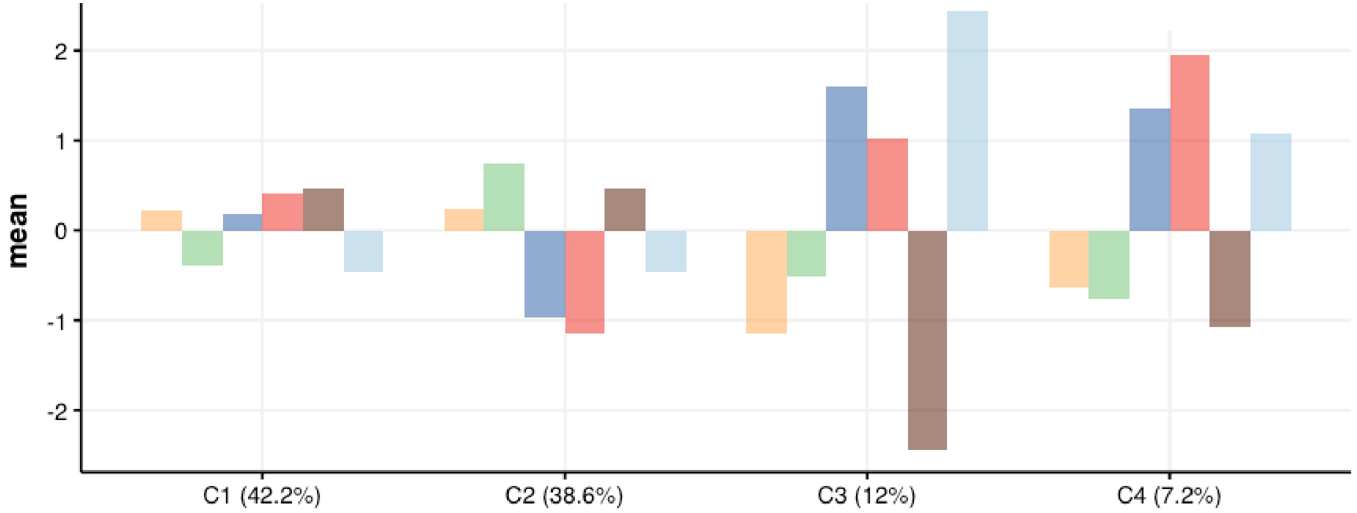
Multiple Correspondence Analysis (MCA)



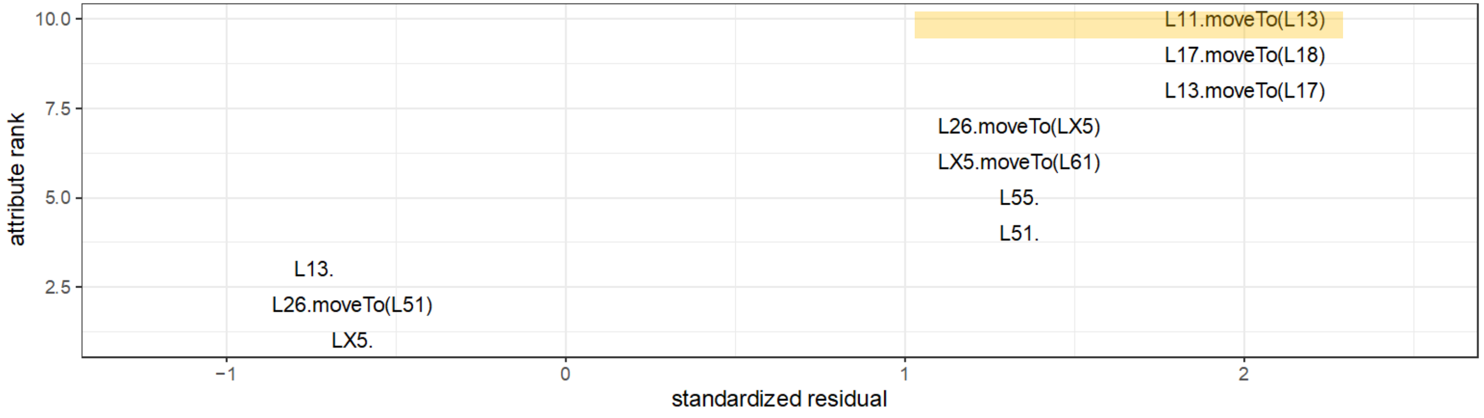
2
1
0
-1
-2
mean
C1 (42.2%)
“the balanced”
C2 (38.6%)
“the fast”
C3 (12%)
“the intrusiveness-avoidant”
C4 (7.2%)
“the moderately intrusiveness-avoidant”
Clusters
Variables
w_collision
w_travelTime
w_intrusiveness
SumTravelTime
SumIntrusiveness
SumCollisions

C1: 42.2% “the balanced”

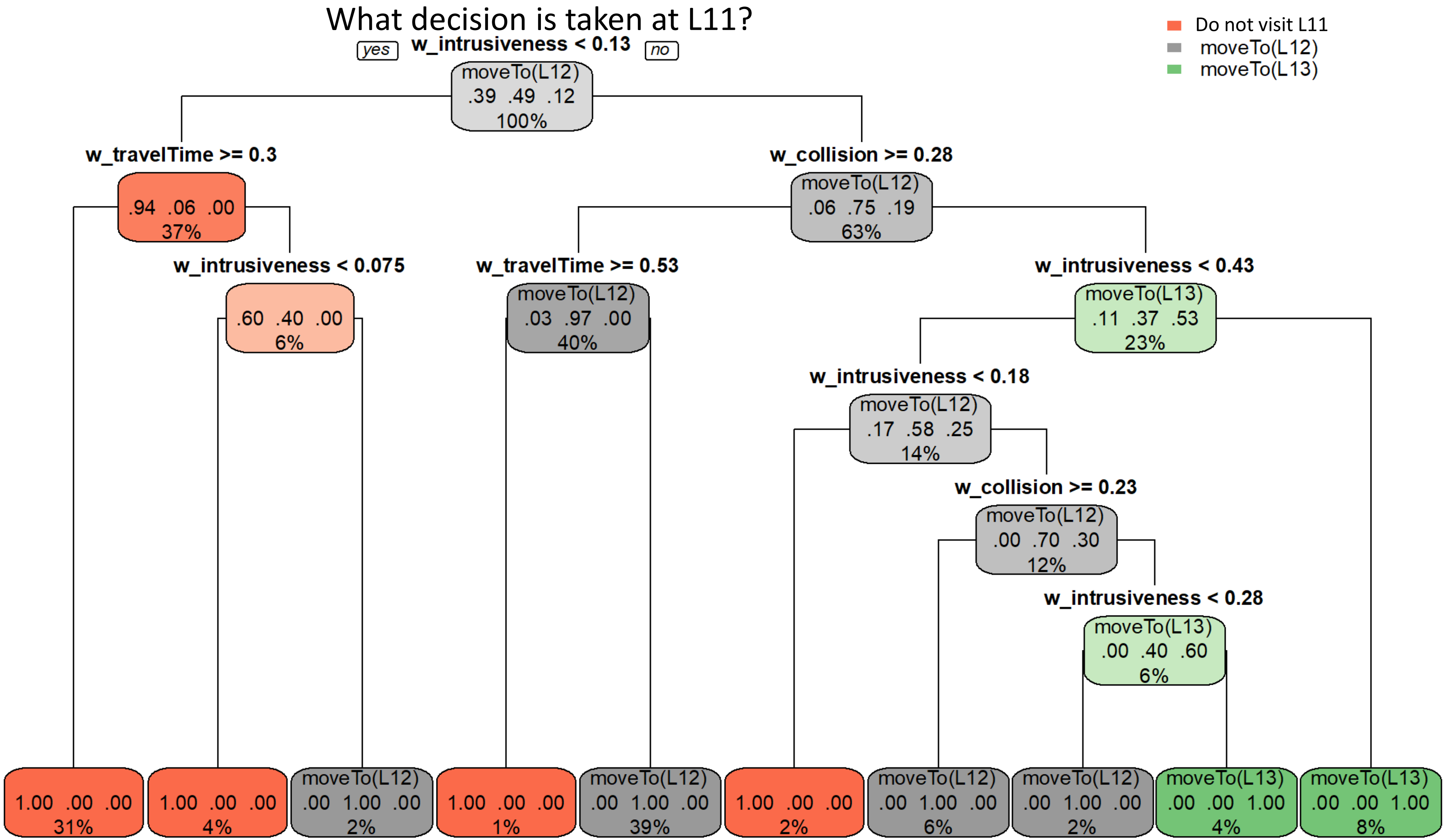




C3: 12% “the intrusiveness-avoidant”



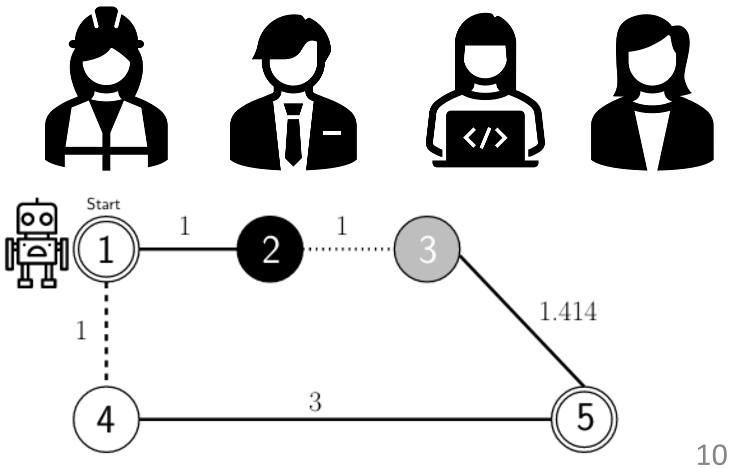
Decision tree learning



Thank you!
David Garlan and Rebekka Wohlrab
garlan@cs.cmu.edu, wohlrab@cmu.edu

Future Work

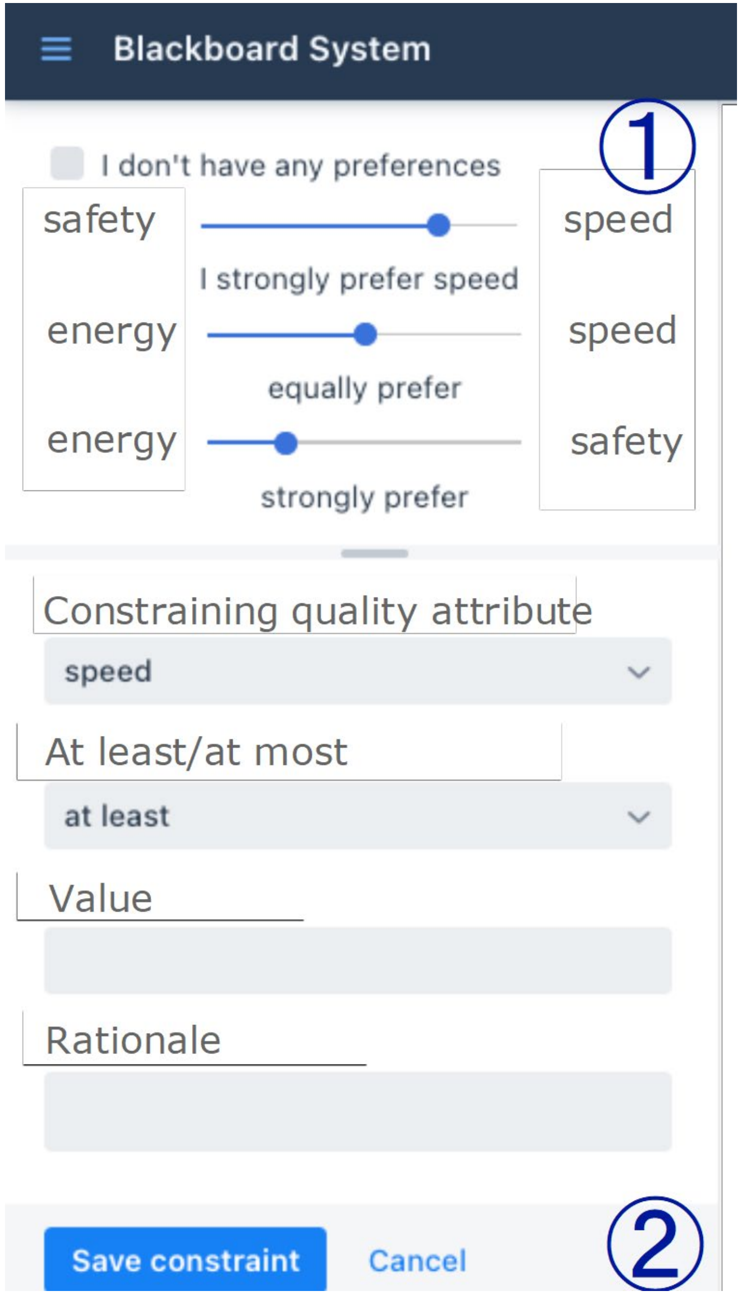
- Apply similar approaches to automated support for cybersecurity
- Create techniques to provide natural language explanations to human stakeholders
- Develop decision support mechanisms to enable humans to ensure that the generated plans meet their requirements



References

- [1] Sukkerd, R., Simmons, R., & Garlan, D. (2020). Tradeoff-focused contrastive explanation for MDP planning. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1041-1048). IEEE.
- [2] Wohlrab, R., Cámara, J., Garlan, D. & Schmerl, B. Explaining Quality Attribute Tradeoffs in Automated Planning for Self-Adaptive Systems. In submission.
- [3] Wohlrab, R., & Garlan, D. A Negotiation Support System for Defining Utility Functions for Multi-Stakeholder Self-Adaptive Systems. Accepted to Requirements Engineering.
- [4] Cámara, J., Silva, M., Garlan, D., & Schmerl, B. (2021, September). Explaining Architectural Design Tradeoff Spaces: A Machine Learning Approach. In *European Conference on Software Architecture* (pp. 49-65). Springer, Cham.

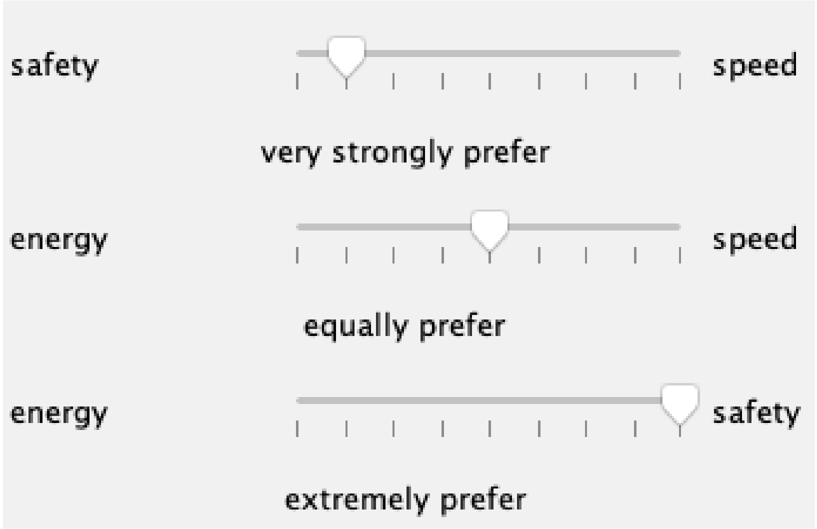
Backup Slides



- Analytic Hierarchy Process

	Safety	Speed	Energy Consumption
Safety	1	7	9
Speed	$\frac{1}{7}$	1	1
Energy Cons.	$\frac{1}{9}$	1	1

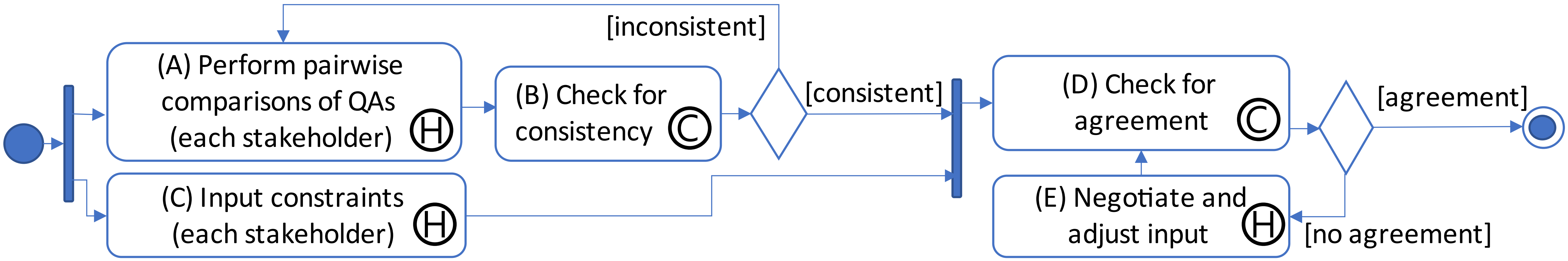
- Pairwise comparison of QAs
- Creation of a reciprocal matrix

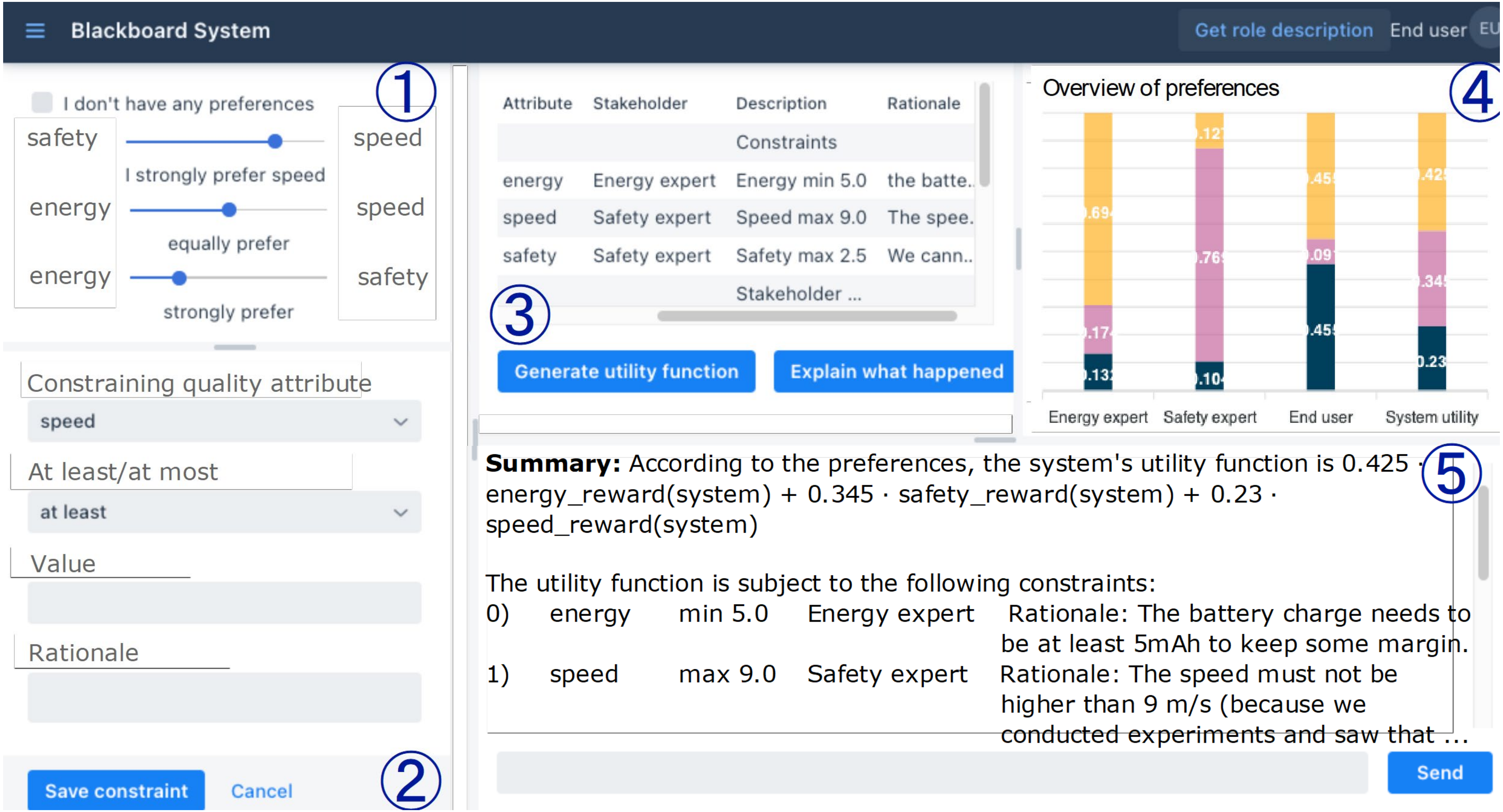


- Normalized principal eigenvector of the matrix A represents the relative priorities of the QAs

utility(plan) = 0.8·utility_speed(cost_speed(plan))+0.1·utility_safety(cost_safety(plan)) +0.1·utility_intrusiveness(cost_intrusiveness(plan))

Method for Utility Function Definition

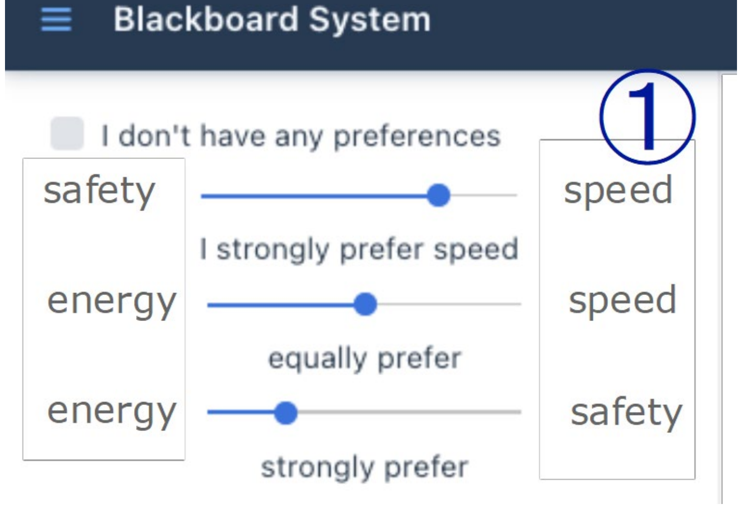


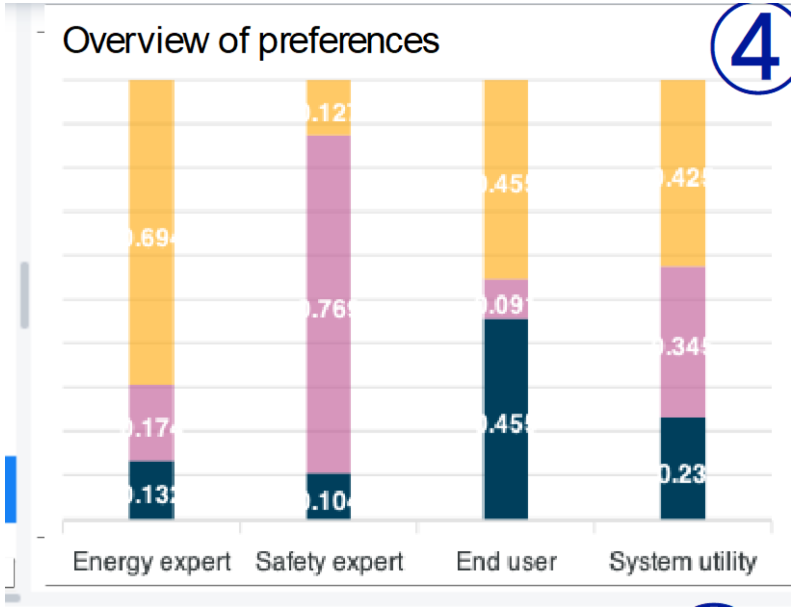

Blackboard System
Get role description
End user
EU
1
I don't have any preferences
safety
speed
I strongly prefer speed
energy
speed
equally prefer
energy
safety
strongly prefer
Constraining quality attribute
speed
At least/at most
at least
Value
Rationale
Save constraint
Cancel
2
Attribute
Stakeholder
Description
Rationale
Constraints
energy
Energy expert
Energy min 5.0
the batte..
speed
Safety expert
Speed max 9.0
The spee..
safety
Safety expert
Safety max 2.5
We cann..
Stakeholder ...
3
Generate utility function
Explain what happened
Overview of preferences
4
Energy expert
Safety expert
End user
System utility
5
Summary: According to the preferences, the system's utility function is 0.425 · energy_reward(system) + 0.345 · safety_reward(system) + 0.23 · speed_reward(system)
The utility function is subject to the following constraints:
0) energy min 5.0 Energy expert Rationale: The battery charge needs to be at least 5mAh to keep some margin.
1) speed max 9.0 Safety expert Rationale: The speed must not be higher than 9 m/s (because we conducted experiments and saw that ...
Send

Concordance of preferences

To reach a consensus, you need to align your preferences.

- Option 1) @End user: To reach a concordant solution, it is enough if you lower the top slider and indicate that you strongly prefer speed over safety. If you do that, you slightly increase your ranking of safety, which is more in line with the others' preferences.
- Option 2) You can also convince the safety expert to lower their preference for safety. If the safety expert prefers safety as much as energy or speed, your preferences are concordant.
- Option 3) You can also convince the energy expert to lower their preference for energy. If the energy expert prefers energy as much as safety or speed, your preferences are concordant. Write in the chat and negotiate with other stakeholders.





Your speed constraint (at least 2.0) is in conflict with the safety expert's constraint (at most 1.0).
It is impossible to state that speed should be both at least 2.0 and at most 1.0.

Safety expert's rationale: The speed should not be higher than 1 m/s (because we conducted experiments and saw that the system would be unsafe otherwise).

End user's rationale: so that the robot can meet its deadlines.

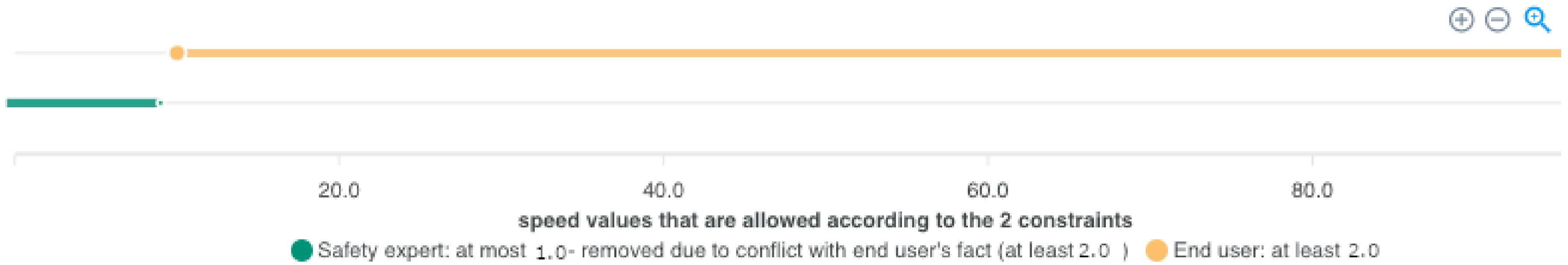
Your authority level for speed is high (2), whereas the safety expert's authority level is the default value (1).

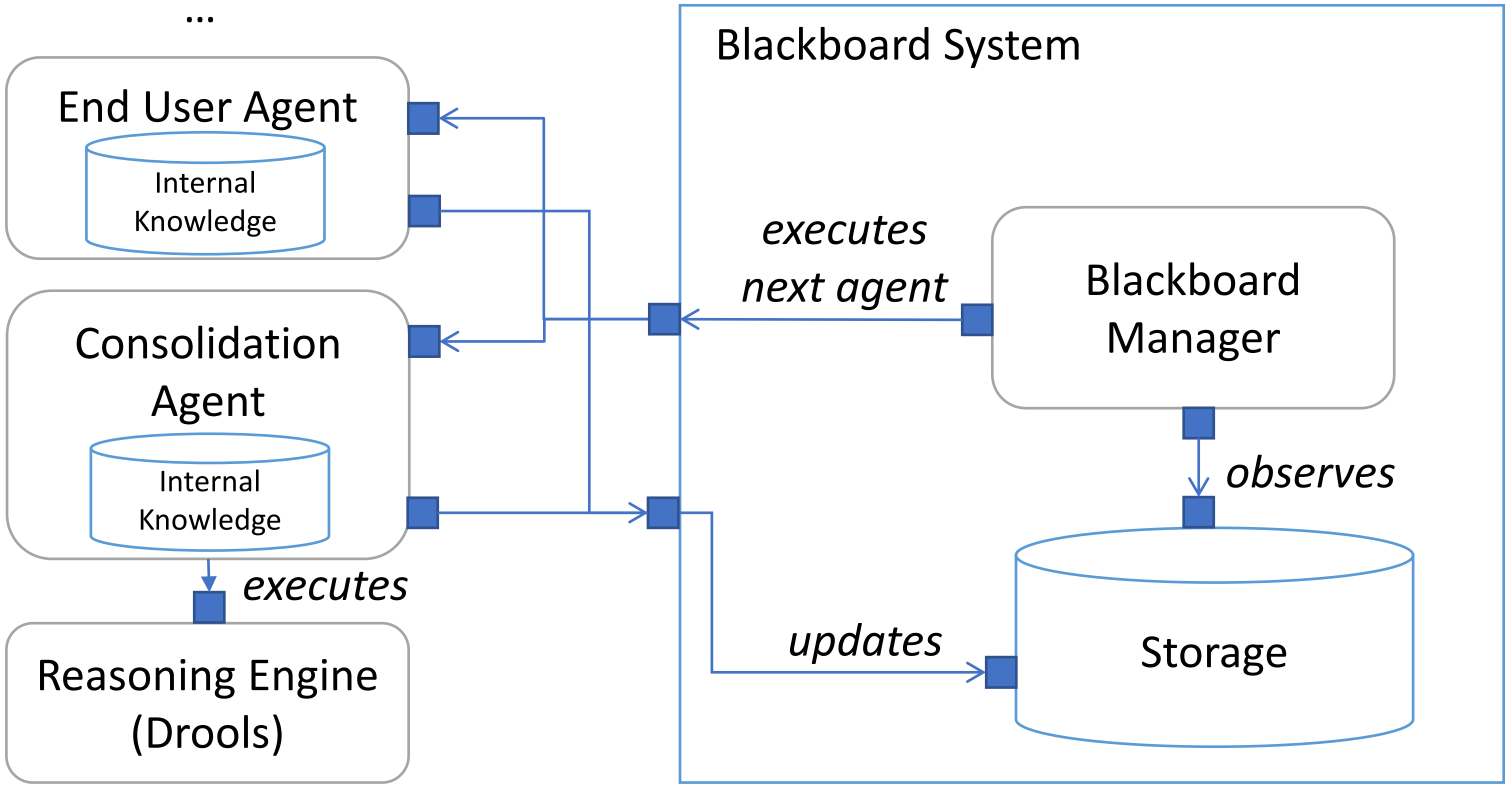
Drop my constraint | Decide based on authority levels (keep my constraint) | Keep both constraints and (re-)negotiate

End user: speed at least 2.0 - This means that all speed values including and above 2.0 satisfy the constraint.

Facts that were removed due to a conflict with this constraint:

1. Constraint: Safety expert: speed at most 1.0
   - Reason: You can't have both at least 2.0 and at most 1.0. The constraint (at most 1.0) was removed.





...
End User Agent
Internal Knowledge
Consolidation Agent
Internal Knowledge
executes
Reasoning Engine (Drools)
Blackboard System
executes next agent
Blackboard Manager
observes
updates
Storage