THuMP project — Enhancing trust in Human-Machine Partnerships

ExplAIn 2021

Gerard Canal (gerard.canal@kcl.ac.uk)
King's College London

08th July 2021

The THuMP project

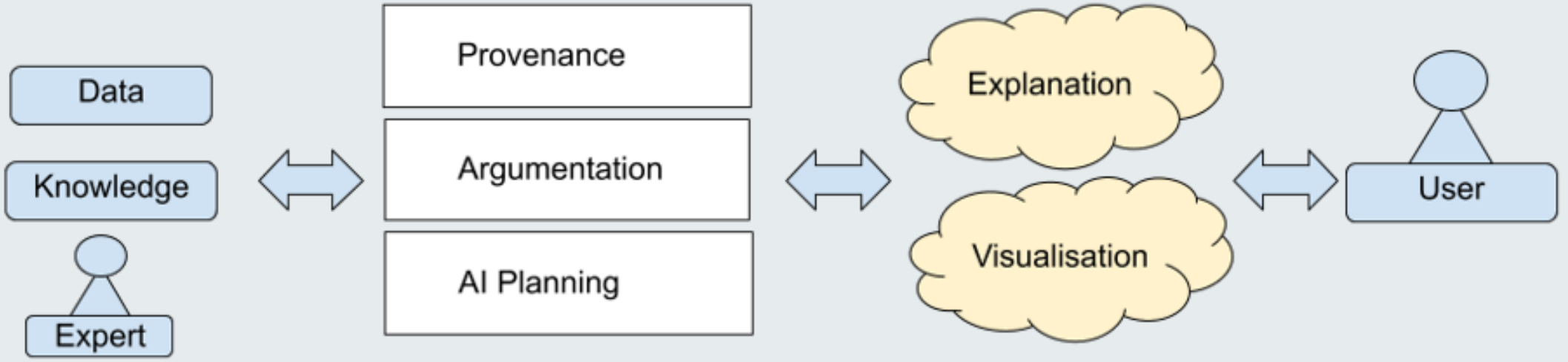
THuMP | trust in human-machine partnerships

EPSRC

Automatic decisions computed to be near-optimal based on a given model may be hard to trust if the system behaves in ways we do not expect or understand.

Thus, users need to understand **what** is the sytem trying to achieve, and **why**.

We believe this trust will come from a multidisciplinary approach. We look at multi-source, multi-modal explanations.



AI planning-based explanations

THuMP | trust in human-machine partnerships

EPSRC

Defining a decision-making problem as a planning problem introduces model and structure that eases the generation of explanations.

Planning is currently applied to many different areas, from robotics to logistics.

Basing decisions on a model allows to reason over it and show the process to the user, presenting the **reasons behind the actions** of the system and **confronting available options**.

AI planning-based explanations

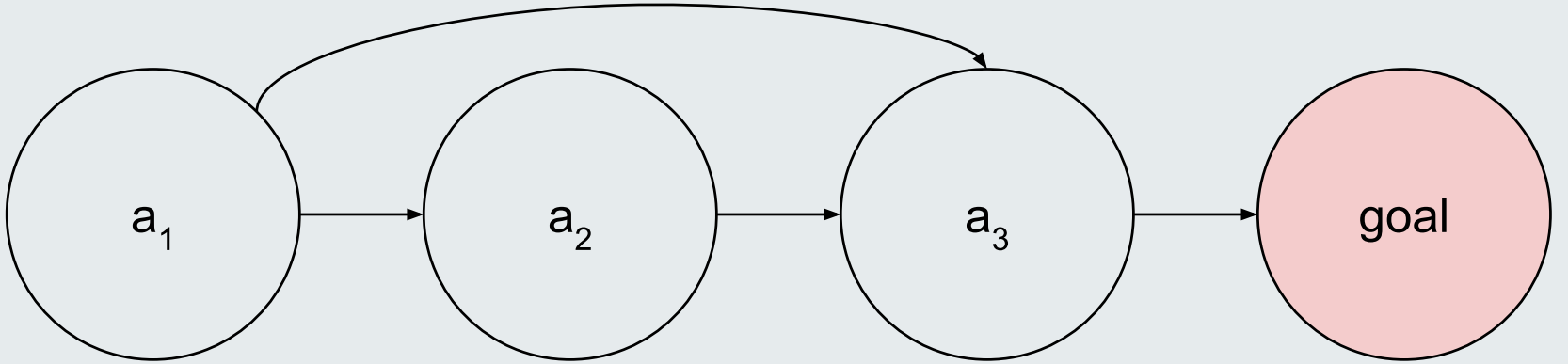
THuMP | trust in human-machine partnerships

EPSRC

Contrastive explanations confront the user's mental model with the planner's output.

In order to show reasons driving actions, we propose plan verbalization.

- We get a plan as an input, and we use the model to compute causal chains from the goals to the initial states (finding enabler actions at each step).
- We summarize the plan, joining intermediate actions when needed.
- We join actions that joinly enable the next action in the plan and keep those that enable later actions in the plan.
- We finally verbalize those actions making the causal information explicit using semantic tags.



AI planning verbalization examples

THuMP | trust in human-machine partnerships

EPSRC

Example of semantic tags:

```
; verb = go / travel / move
; subject = ?v
; prep = from the ?a
; prep = to the ?b / towards the ?b !
(:durative-action goto_waypoint
 :parameters (?v - robot ?a ?b - waypoint)
```

Examples of verbalized actions in a domain with two robots acting in an office environment

Tomo will locate the manager, which will allow me to **later** request the manager at the kitchen corridor and me to hand post2 to the manager at the kitchen corridor.

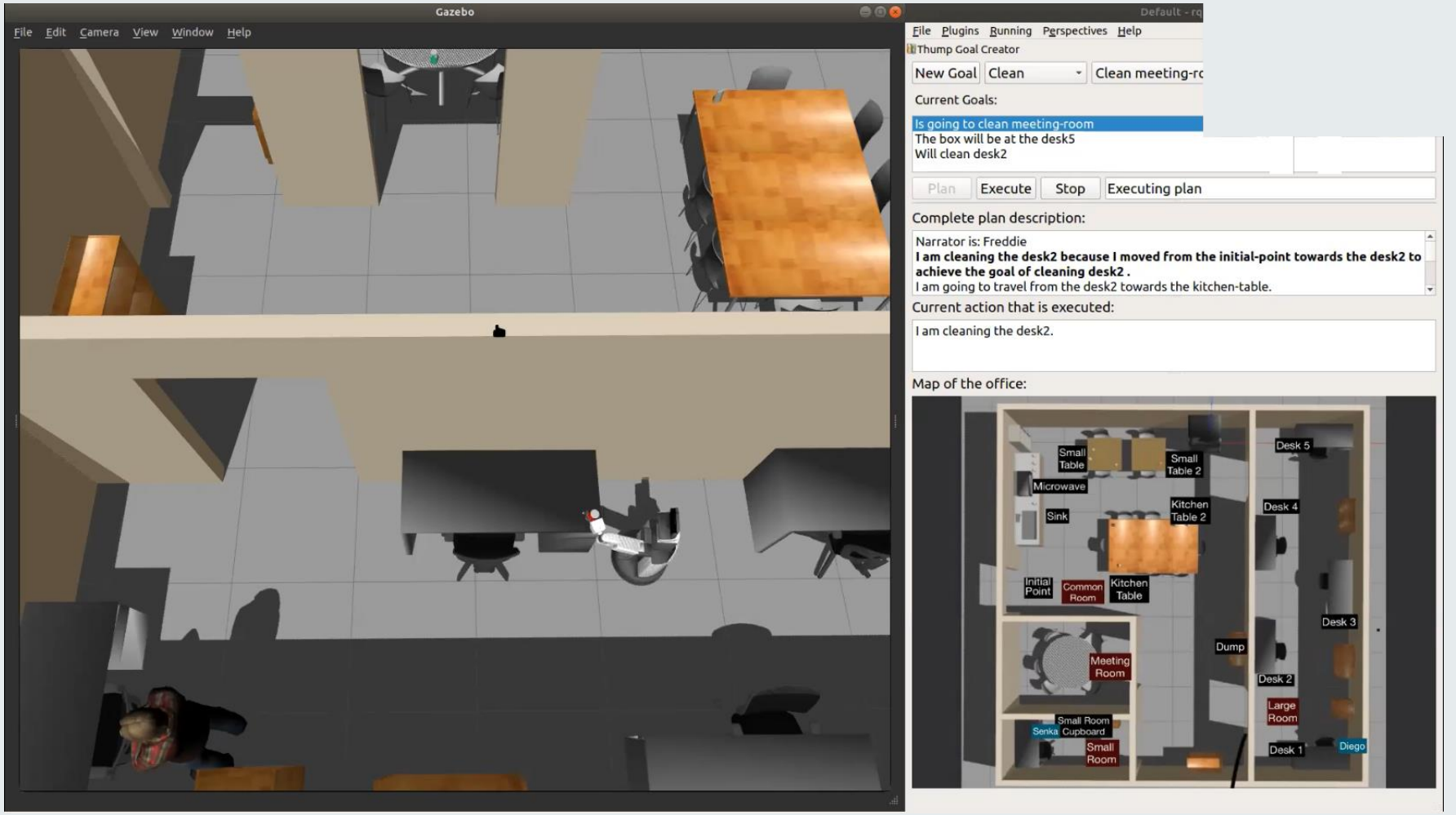
I will travel from the kitchen shelf towards the kitchen counter (via coffee table 1 and kitchen corridor) so I can leave the paper at the kitchen counter to achieve the goal of the paper being at the kitchen counter.

Ongoing user study

THuMP | trust in human-machine partnerships

EPSRC

We are currently running a user study where we evaluate the verbalizations in action, when the users are in charge of defining the tasks of the robot in a collaborative manner and seeing the results online.



THANK YOU FOR YOUR ATTENTION!

THuMP | trust in human-machine partnerships

KING'S College LONDON

Gerard Canal
King's College London