Balancing Explicability and Explanations for Human-Aware Planning

Tathagata Chakraborti, Sarath Sreedharan and Subbarao Kambhampati

SAPIENZA
UNIVERSITÀ DI ROMA

Vincenzo Colella

Introduction

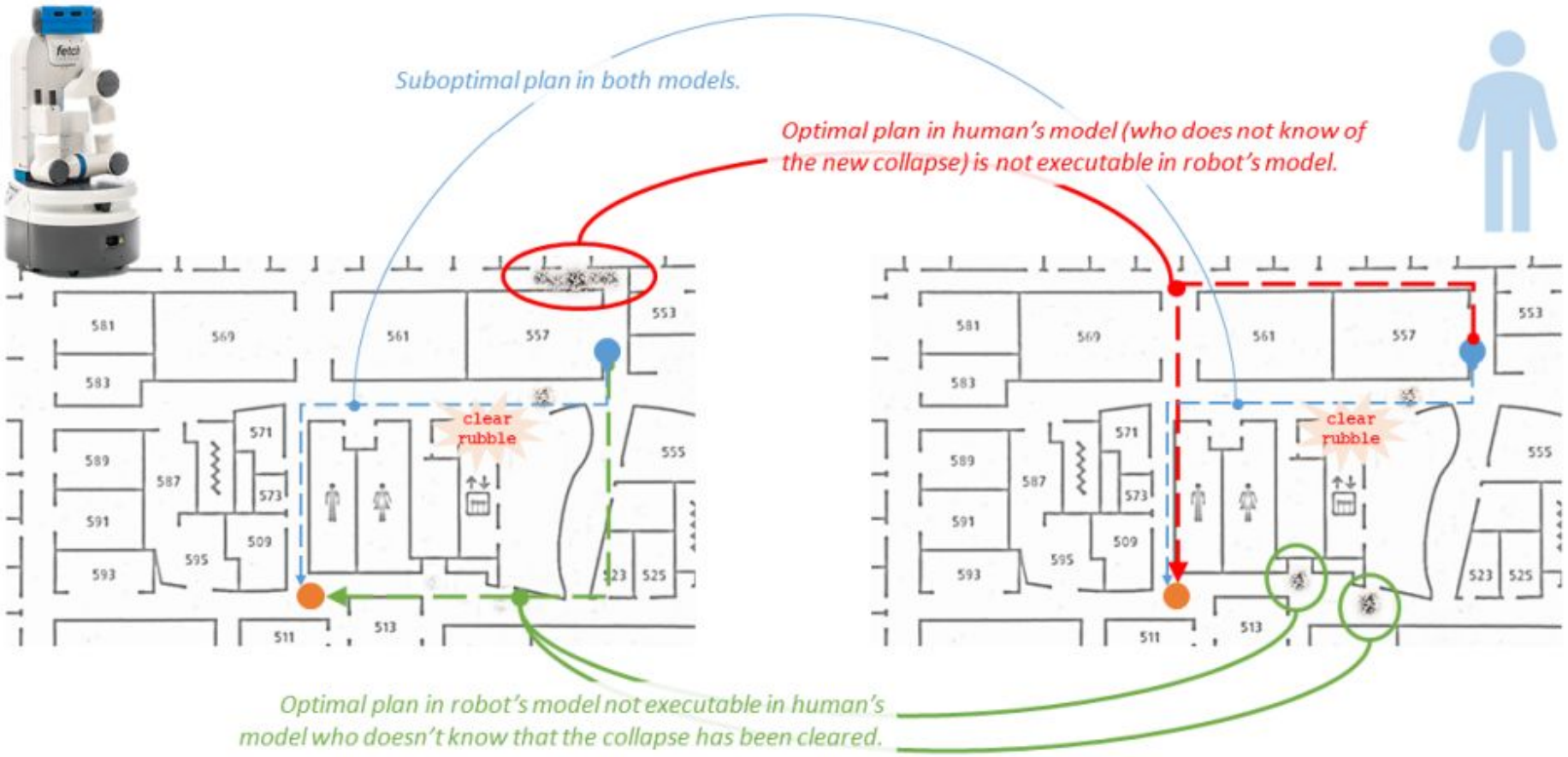
- When a robot is working under human supervision, the mental model of the task that the robot will create (M_R) often differs from the model that the human thinks it has (M_h).

WHY ?

- This divergence may appear while a robot is working on its task and notices that a path has collapsed, and may update its mental model.
- It may also happen that a path that wasn't available is cleared without the human noticing it.

➔ In this paper we will analyze the MEGA algorithm, that manages the explicability-explanation trade-off whenever a robot has a different model than its human agent during a task.

Explicability - Explanation Trade-off



MEGA Algorithm

- The algorithm proposed in the paper manages the strategy that the robot will follow when in this circumstance.
- The robot must choose between:
  - ➔ The most explicable plan
    - ◆ The closest to the human expectation, which needs the least amount of explanations
  - ➔ The less explicable plan
    - ◆ The furthest from the expectation, that needs many explications (=> but better)
- MEGA is parametrized by an alpha value.

Explicability - Explanation Trade-off Parametrized by α

LOW α

HIGH α

MEGA Output

$$
\begin{aligned}
&(1)\ \ \delta_{\mathcal{M}^R}(\mathcal{I}^R, \pi) \models \mathcal{G}^R; \\
&(2)\ \ \bar{\mathcal{M}}_h^R \longleftarrow \mathcal{M}_h^R + \mathcal{E}; \\
&(3)\ \ C(\pi, \bar{\mathcal{M}}_h^R) = C^*_{\bar{\mathcal{M}}_h^R}; \text{ and} \\
&(4)\ \ \pi = \arg\min_\pi \ \{\, |\mathcal{E}| \ + \ \alpha \times |\, C(\pi, \mathcal{M}^R) - C^*_{\mathcal{M}^R} \,|\, \}.
\end{aligned}
$$

- The output of MEGA Algorithm is a plan π and an explanation Ɛ such that the plan is executable in robot's model **(1)** and the explanation in the form of model updates **(2)**, optimal in the updated human model **(3)**, while the cost (length) of explanations and the cost of deviation from optimality in its own model to be explicable is traded off according to a constant α **(4)**.

MEGA Experiments

```
(fill-shot shot2 ingredient2 left right dispenser2)
(pour-shot-to-used-shaker shot2 ingredient3 shaker1 left)
(refill-shot shot2 ingredient3 left right dispenser3)
(pour-shot-to-used-shaker shot2 ingredient3 shaker1 left)
(leave left shot2)
(grasp left shaker1)
```

Fig.1: Human Expectation

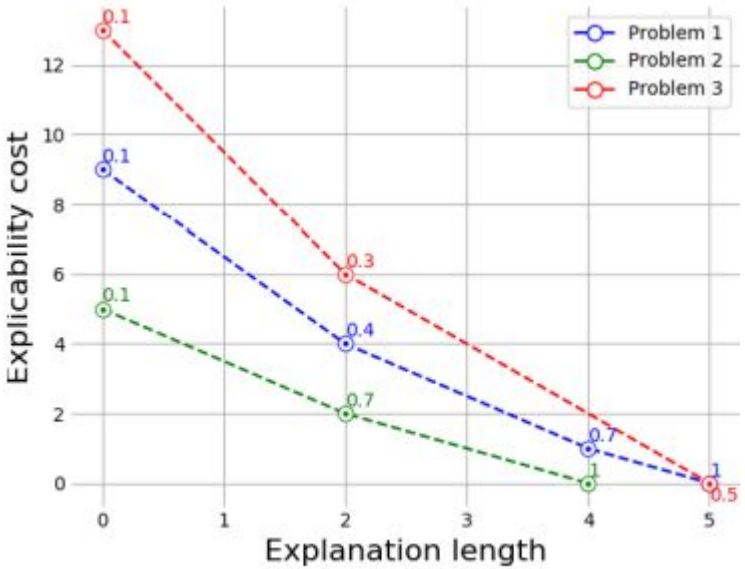


Fig.2: Trade-off

Conclusions

- We saw how an agent can be human-aware by balancing the cost of departure from optimality (in order to conform to human expectations) versus the cost of explaining away causes of perceived suboptimality.

- It is well known how *humans make better decisions when they have to explain [Mercier and Sperber, 2011].*

- In this work, in being able to reason about the explainability of its decisions, an AI planning agent is similarly able to make better decisions by explicitly considering the implications of its behavior on the human mental model.

Balancing Explicability and Explanations for Human-Aware Planning

Tathagata Chakraborti, Sarath Sreedharan and Subbarao Kambhampati

SAPIENZA
UNIVERSITÀ DI ROMA

Thank you !