PLEASE: Generating Personalized Explanations in Human-Aware Planning

Supplement

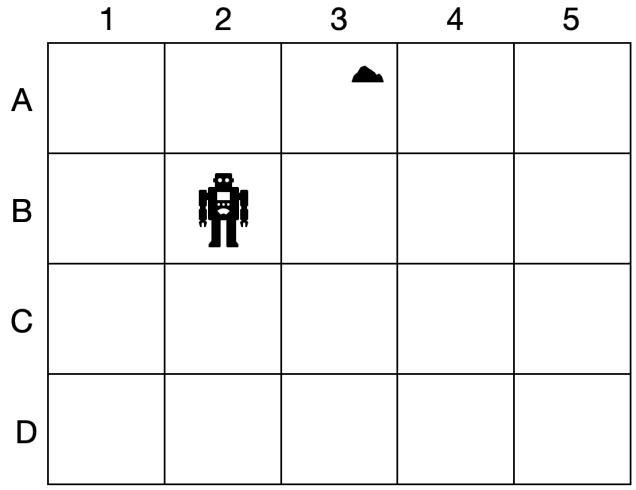


Figure 1: The 5x4 grid shown to all participants.

1 Human User Study

Figure 1 depicts the 5x4 grid that was shown to all participants in the study. Recall that the participants were informed that the robot moved from location B2 to A3 and then made a decision for which an explanation was generated.

1.1 Explanation Generation

The explanations for the participants were generated from the following knowledge base, i.e., the robot's knowledge base:

$$KB_a = \{\Phi, \Omega, \neg\Phi \vee \neg\Omega \vee \Lambda, \neg\Lambda \vee \Psi, \neg\Psi \vee C\}$$

with vocabulary

$$\begin{aligned}\mathcal{V}_{KB_a} = \{&\Phi : \text{rock-at(A3)}\\ &\Lambda : \text{sample-rock(A3)}\\ &\Omega : \text{handempty}\\ &\Psi : \text{have-analysis(rock)}\\ &C : \text{communicate-data(rock)}\}\end{aligned}$$

The vocabulary $\mathcal{V}_{KB_a}$ describes the meaning of the symbols used in the formulae of KB_a. The explanandum for the study was $\varphi = C$, where $KB_a \models C$.

The participants were divided into three vocabulary pairs, with each pair consisting of a treatment group and a control group. The pairs received the following vocabularies:

$$\begin{aligned}&\textit{Pair 1: } \mathcal{V}_{h_1} = \{\Phi\}\\ &\textit{Pair 2: } \mathcal{V}_{h_2} = \{\Phi, \Lambda\}\\ &\textit{Pair 3: } \mathcal{V}_{h_2} = \{\Phi, \Lambda, \Omega\}\end{aligned}$$

The treatment groups in each pair received a personalized explanation w.r.t. their vocabulary, where we used an upper bound $\mathcal{UB} = 4$. Specifically, the personalized explanations for each treatment group are as follows:

Treatment group 1:

$$\begin{aligned}\epsilon_{t_1} &= \mathcal{F}(KB_a, \{\Lambda, \Omega, \Psi\})\\ &= \{\Phi, \neg\Phi \vee C\}\end{aligned}$$

Treatment group 2:

$$\begin{aligned}\epsilon_{t_2} &= \mathcal{F}(KB_a, \{\Omega, \Psi\})\\ &= \{\Phi, \neg\Phi \vee \Lambda, \neg\Lambda \vee C\}\end{aligned}$$

Treatment group 3:

$$\begin{aligned}\epsilon_{t_3} &= \mathcal{F}(KB_a, \{\Psi\})\\ &= \{\Phi, \Omega, \neg\Phi \vee \neg\Omega \vee \Lambda, \neg\Lambda \vee C\}\end{aligned}$$

The control groups in each pair received the same generic explanation, i.e.:

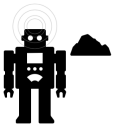
Control groups (for all pairs):

$$\begin{aligned}\epsilon_c &= \mathcal{F}(KB_a, \{\emptyset\})\\ &= \{\Phi, \Omega, \neg\Phi \vee \neg\Omega \vee \Lambda, \neg\Lambda \vee \Psi, \neg\Psi \vee C\}\end{aligned}$$

Figure 2 shows the natural language translation of the explanations shown to the groups in each pair.

1.2 Results

We now present all of the results and analysis thereof. Upon seeing the explanations, the groups were asked to evaluate the explanations by answering the following questions:

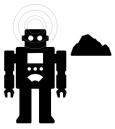


Φ
rock-at(A3)

Treatment Group Explanation:
You told me to move to location A3, but because of Φ I communicated that data to the station.

Control Group Explanation:
You told me to move to location A3, but because of Φ and Ω, I executed Λ. Then, because I had done Λ, I resulted in doing Ψ. And because I had done Ψ, I communicated that data to the station.

(a)

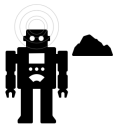


Φ	Λ
rock-at(A3)	sample-rock(rock,A3)

Treatment Group Explanation:
You told me to move to location A3, but because of Φ, I executed Λ. Then, because I had done Λ, I communicated that data to the station.

Control Group Explanation:
You told me to move to location A3, but because of Φ and Ω, I executed Λ. Then, because I had done Λ, I resulted in doing Ψ. And because I had done Ψ, I communicated that data to the station.

(b)



Φ	Ω	Λ
rock-at(A3)	handempty	sample-rock(rock,A3)

Treatment Group Explanation:
You told me to move to location A3, but because of Φ and Ω, I executed Λ. Then, because I had done Λ, I communicated that data to the station.

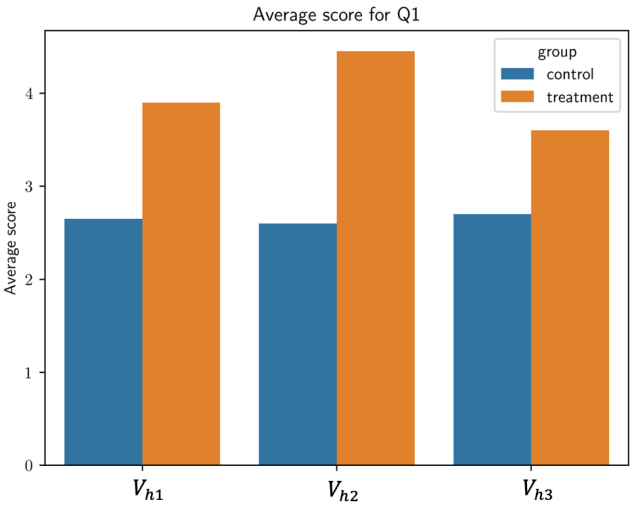
Control Group Explanation:
You told me to move to location A3, but because of Φ and Ω, I executed Λ. Then, because I had done Λ, I resulted in doing Ψ. And because I had done Ψ, I communicated that data to the station.

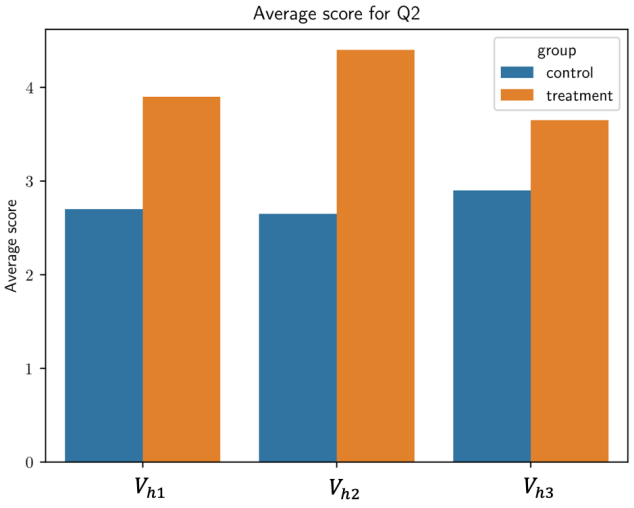
(c)

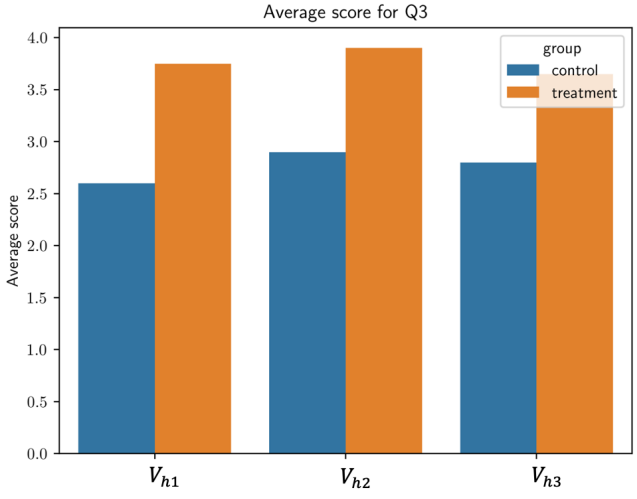
Figure 2: The explanations shown to the groups in each vocabulary pair, where (a) are the explanations for pair 1 $\mathcal{V}_{h1}$, (b) for pair 2 $\mathcal{V}_{h2}$, and (c) for pair 3 $\mathcal{V}_{h3}$

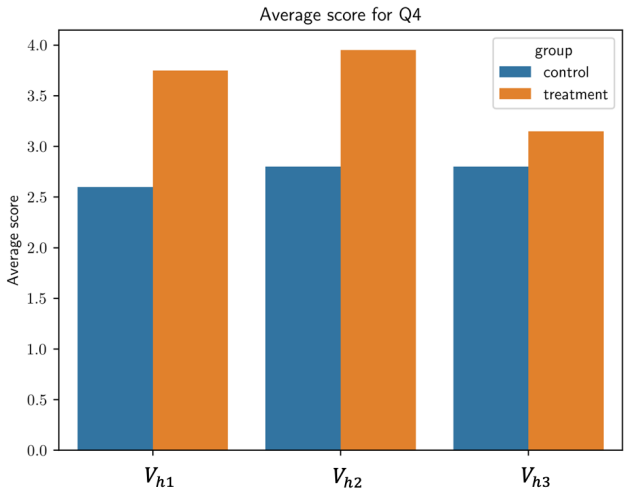
Q1. **The explanation helped me understand the robot's decision to communicate the data.** (Likert-type: 1: *Strongly disagree* - 5 : *Strongly agree*)
Q2. **I am satisfied with the robot's explanation about how it behaved.** (Likert-type: 1: *Strongly disagree* - 5 : *Strongly agree*)
Q3. **I feel that the explanation of how the robot behaved has sufficient detail.** (Likert-type: 1: *Strongly disagree* - 5 : *Strongly agree*)
Q4. **I feel that the explanation of how the robot behaved is complete.** (Likert-type: 1: *Strongly disagree* - 5 : *Strongly agree*)
Q5. **How useful do you find the robot's explanation for understanding its behavior?** (Likert-type: 1: *Not useful at all* - 5 : *Extremely useful*)
Q6. **How confident are you in your understanding of the explanation?** (Likert-type: 1: *Not confident at all* - 5 : *Extremely confident*)
Q7. **How confident are you in your ability to explain the robot's behavior (based on its explanation) to someone else?** (Likert-type: 1 : *Not confident at all* - 5 : *Extremely confident*)
Q8. **Do you think having access to a vocabulary of task-specific terms helped improve your understanding of the explanation?** (Yes, No, Maybe)
Q9. **In future interactions with AI agents, would you prefer personalized explanations or generic ones?** (Personalized explanations, generic explanations)

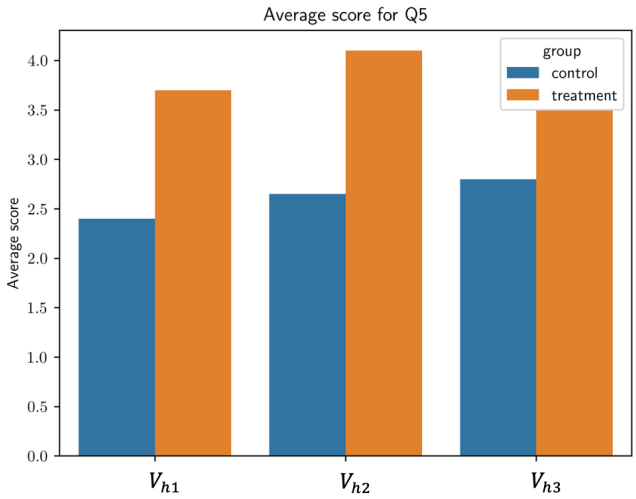
The distributions of each questions are shown in the following figures:

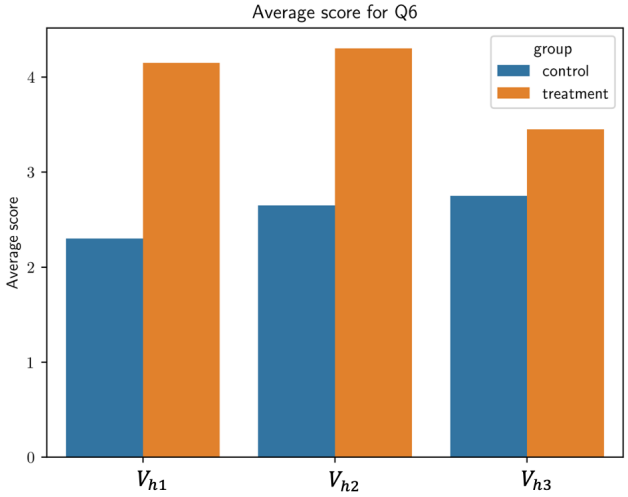


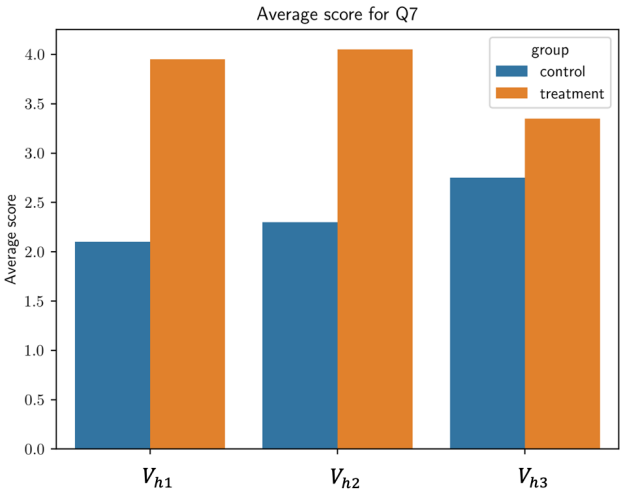


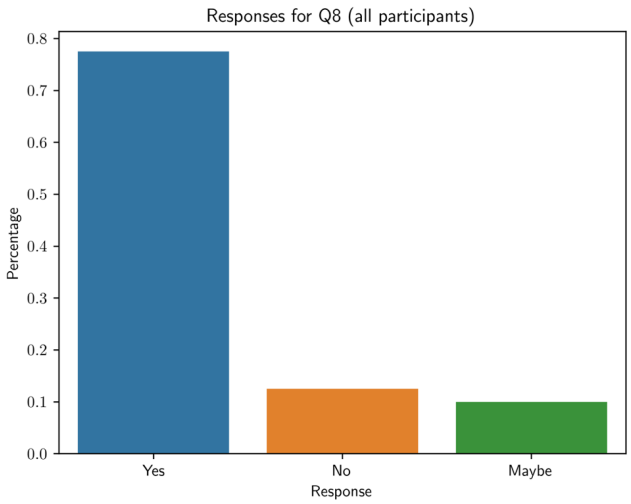


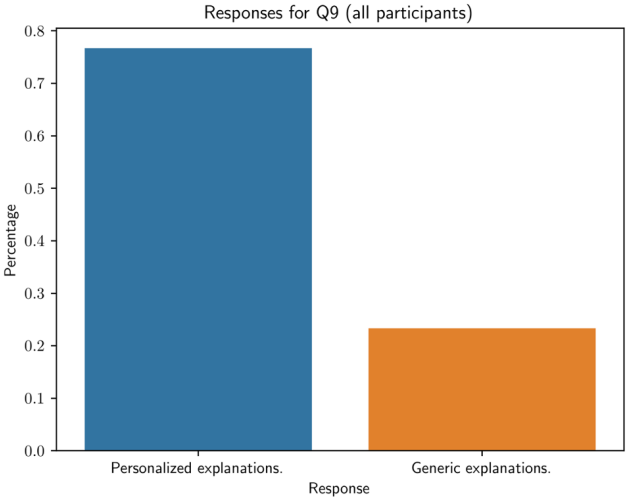

Average score for Q4
group
control
treatment
Average score
V_{h1}
V_{h2}
V_{h3}


Average score for Q5
group
control
treatment
Average score
V_{h1}
V_{h2}
V_{h3}


Average score for Q6
group
control
treatment
Average score
V_{h1}
V_{h2}
V_{h3}


Average score for Q7
group
control
treatment
Average score
V_{h1}
V_{h2}
V_{h3}


Responses for Q8 (all participants)
Percentage
Yes
No
Maybe
Response


Responses for Q9 (all participants)
Percentage
Personalized explanations.
Generic explanations.
Response