Explainable AI for human supervision over firefighting robots

The influence of on-demand explanations on human trust

Elena Negrila E.Negrila@student.tudelft.nl

Supervisor: Ruben Verhagen Responsible Professor: Myrthe Tielman

1. Introduction & Background

- Collaboration between humans and robots in 2D simulated firefighting environment
- Semi-autonomous robot assesses situation and reports to human supervisor
- Search, rescue & extinguish operations done by the robot
- Tasks implying moral decisions are sensitive [1]
- Robot takes a decision if the moral sensitivity less than pre-defined threshold

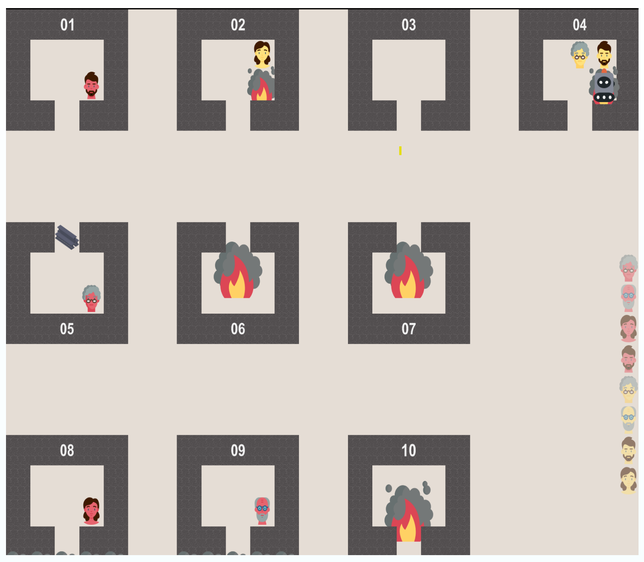

Figure 1: Simulated firefighting nvironment

2. Research Questions

How does the possibility of requesting additional on-demand explanations, compared to receiving the already existing baseline explanations, influence the level of trust of the human supervisor in the robot?

- How do the on-demand explanations differ from the baseline explanations?
- How often do users require on-demand explanations from the robot?
- Does the background of the users impact how frequently they require explanations?
- Is the frequency of the request for additional information correlated to the level of trust in the robot?
- Do the users still need more insight, even after receiving this extra information?

3. Method

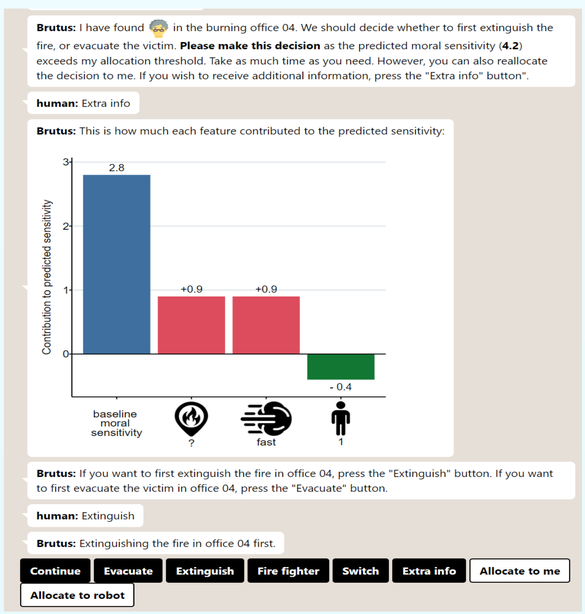

Figure 2: Interface for on-demand condition

- Visual explanations are only displayed only when the implemented '**Extra info**' button is pressed
- **User study** with 40 participants (20 interacted with the baseline condition, 20 with the on-demand)
- **Capacity trust, moral trust, explainable AI satisfaction, disagreement rate** measured through a pre-defined questionnaire in Qualtrics

4. Results

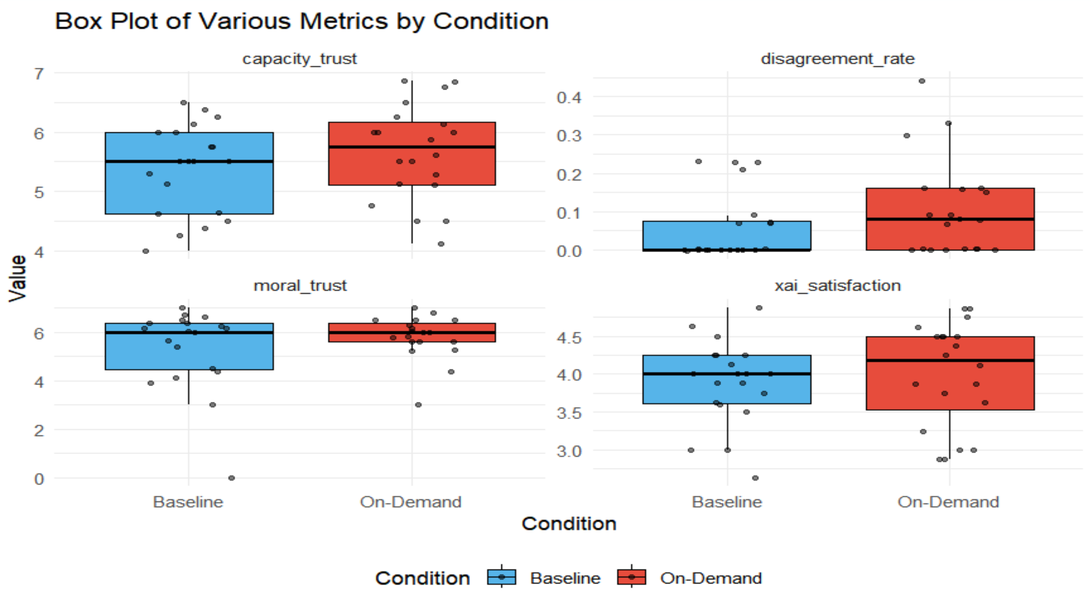

Figure 3: Data distribution in baseline VS on-demand

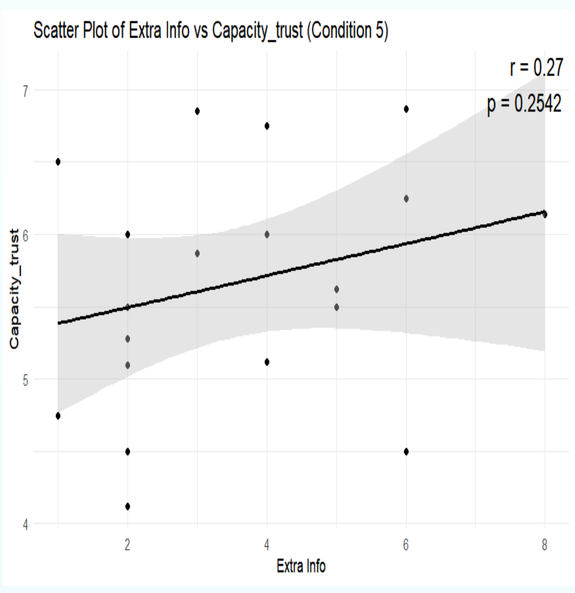

Figure 4: Correlation between on-demand frequency requests and capacity trust

- High mean values for both conditions
- Low correlation between the frequency of request for on-demand explanations and overall trust and satisfaction
- User's background does not impact frequency of requests
- Data analysis on dependent variables resulted in no statistically relevant difference between conditions

5. Discussion & Conclusions

- Consistently high scores of trust in the robot's decision, of satisfaction with the explanations, and low disagreement rate.
- Visual explanations might have caused information overload, reducing their impact.
- The baseline version already had good levels of trust and satisfaction, so the quality of the interaction was not significantly changed.
- Announcing the robot's intended actions in advance could improve the collaboration.
- Setting a higher moral sensitivity threshold could allow more decisions and better trust exploration.
- User studies with more participants could provide more staistical data and reveal differences

References:

[1] J. van der Waa, J. van Diggelen, L. Cavalcante Siebert, M. Neerincx, and C. Jonker, "Allocation of moral decision-making in human-agent teams: A pattern approach," in Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19â24, 2020, Proceedings, Part II, pp. 203–220, Springer International Publishing, 2020.

TU Delft | AI Initiative