Influence of Gradually Abstracting Adaptive Explanations on Human Supervision and Trust in Robots

Author: Elena Ibanez (e.ibanez@student.tudelft.nl)
Supervisor: Ruben Verhagen
Responsible Professor: Myrthe Tielman

1. Background

- **Human - Agent Teamwork (HAT):** AI developments enable agents to collaborate with humans, sharing objectives and responsibilities.
- **Explainable AI (XAI):** refers to methods make AI decisions understandable to humans, enhancing trust and collaboration.
- **Adaptive Explanations:** tailor information to the user's knowledge level, improving communication efficiency and user satisfaction in human-agent interactions.

2. Research Question

RQ: *How do adaptive explanations that become more abstract over time influence human supervision over and trust in the robot?*

3. Scenario & Task

- **Environment**: A dynamic task allocation system in MATRX simulated a 2D firefighting environment with 11 victims needing rescue (see Figure 1).
- **Task**: The robot made decisions based on predicted moral sensitivity, deferring to the human supervisor if the threshold was exceeded.
- **Agent**: Brutus, a firefighting agent, performed search and rescue operations in collaboration with a human supervisor.

4. Measures

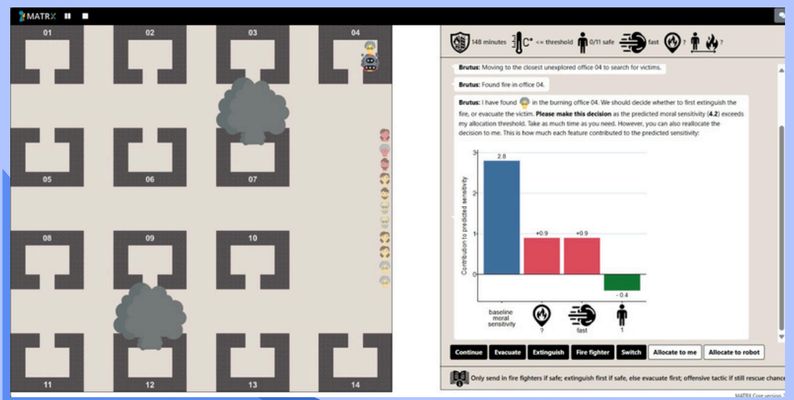
a) Dependent Variables

- Capacity Trust
- Moral Trust
- Disagreement Rate

b) Control Variables

- Demographic Variables
- Gaming Experience
- Risk Propensity
- Trust Propensity
- Utilitarianism

Figure 1. Screenshot of the environment



5. Adaptive Explanations Design

- Motivation: Experienced participants need fewer detailed explanations, reducing information load.
- Initial phases require granular explanations, but more abstract ones become adequate as familiarity increases (see Example).
- The adaptive strategy involves evolving plots and explanations through stages, maintaining essential information flow (See Figure 2)

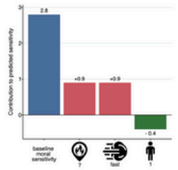


Figure 1: **Early Stage, plot with a full explanation.**

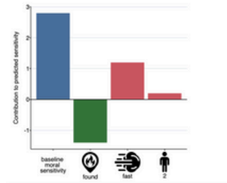


Figure 2: **Intermediate Stage, somewhat detailed.**

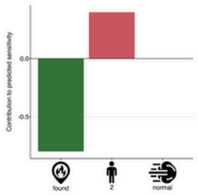


Figure 3: **Abstract Stage, less detail.**

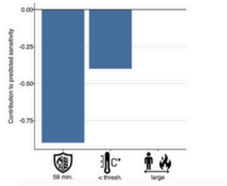


Figure 4: **Very Abstract Stage, minimal detail.**

Figure 2. Screenshots of the plots

Example:

Number of Occurrence (2) -> *Active for {deployment_time} minutes. Continue or switch to defense? Decision needed due to sensitivity ({sensitivity}). Take your time or assign it to me. Feature contributions: (plot)*

Number of Occurrence (6) -> *Continue or switch to offense? Decision needed due to sensitivity({sensitivity}). Contributions: (plot)*

6. User Study

- Involved 40 participants, with 20 assigned to a baseline/non-adaptive scenario and 20 to an adaptive scenario.
- Participants provided informed consent and completed the survey.
- All survey responses were collected using Qualtrics.

7. Results

- Statistical tests on the four dependent variables (capacity trust, moral trust, XAI satisfaction, and disagreement rate) show no significant differences between baseline/non-adaptive and adaptive explanations (see Figure 3).
- This indicates that the two conditions do not notably impact these measures.

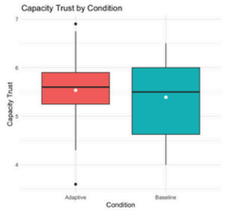


Figure 5: Capacity Trust

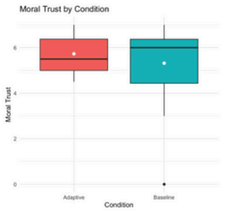


Figure 6: Moral Trust

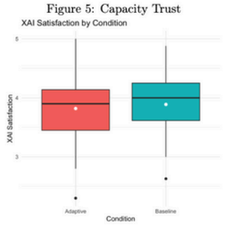


Figure 7: XAI Satisfaction

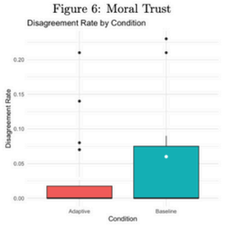


Figure 8: Disagreement Rate

Figure 3: Screenshots of the Boxplots.

8. Discussion & Conclusion

- No significant differences found between baseline/non-adaptive and adaptive explanations.
- Indication of high trust and satisfaction across both conditions suggest participants perceived the robot as trustworthy and were satisfied with the explanations.

& Future research should explore more diverse participant groups, alternative adaptive explanations, and long-term effects to better understand their potential benefits.