Influence of Global Explanations on Human Supervision & Trust in Agent

Explainable AI for human supervision over firefighting robots

Dafni Pandeva

Delft University of Technology, D.V.Pandeva@student.tudelft.nl

Introduction

- AI agent integration becoming more widespread
- Effective human-agent collaboration is essential in critical domains
- Challenges:
- Allocation of moral decision-making tasks
- Maintaining meaningful human control
- Usage of explanations - combat the “black-box” notion and lead to better collaboration
- Knowledge gap on specific types of explanation
- Main focus : **Global Explanations (GEs)**

Scenario & Objective

- Firefighting agent performing search & rescue operations in collaboration with human
- Morally sensitive situations above a certain threshold are assigned to human
- Unknown number of victims and 6 situational features to mimic reality

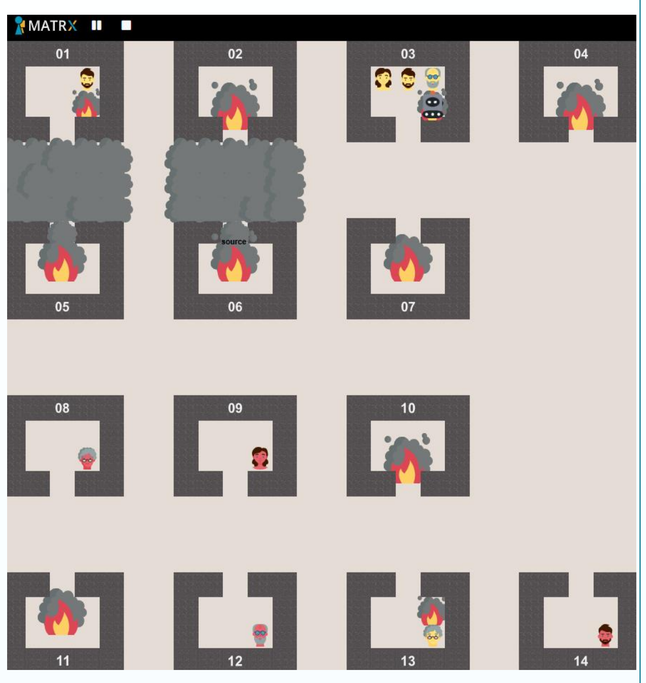


Figure 1: Environment of the model used in the scenario

Methodology

- Literature Review
- Global Explanations Prompts Design
- User Study
- Data Analysis
- Conclusion

Key Characteristics of GEs

- Identify and describe general patterns or trends given a certain situation
- Provide broad view of the logic behind the model
- Summarize more complex explanations into more understandable representations
- Offer scalability

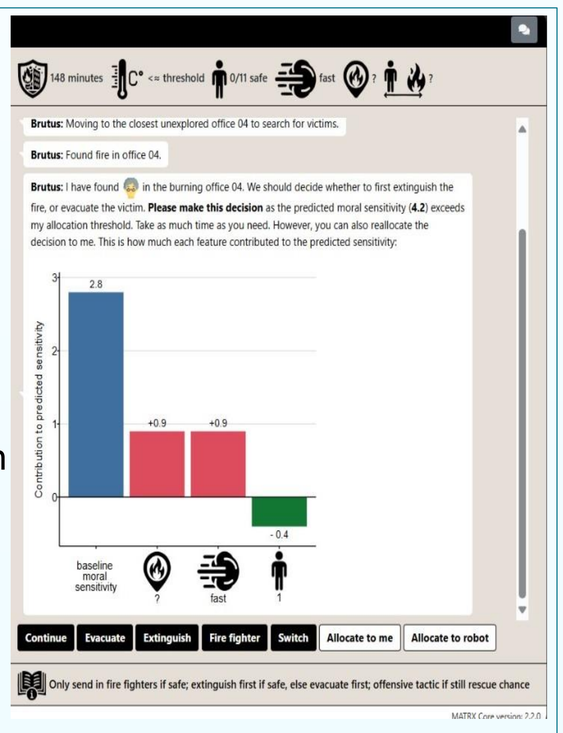


Figure 2: Chat window showing baseline explanation

Global Explanations Design

- Prompt design

– identify main features influencing a decision and their ranges
– analyse the resulting moral sensitivity outcomes and derive general statements
– implement general template structure, e.g.:

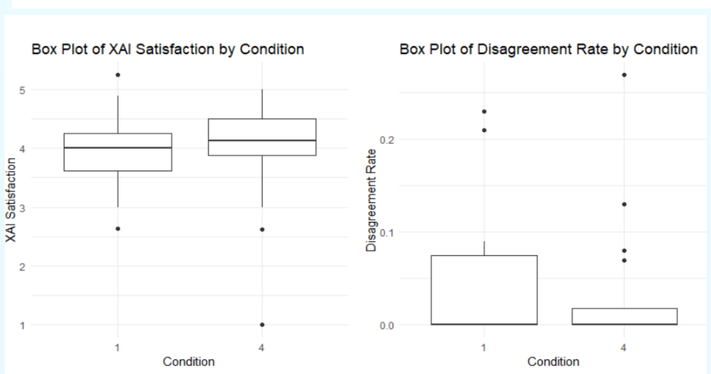
> I have found {victim-type} in office {office-number}. We should decide whether to send in a firefighter to rescue the victim, or if this is too dangerous. I will make this decision as the predicted moral sensitivity {sensitivity-value} is below my allocation threshold. However, you can also reallocate the decision to yourself. {**When the distance is small and the temperature is higher than the threshold, the moral sensitivity is below the allocation threshold in 76.92% of cases.** }

User Study

- Perform user study – 40 participants
- 20 p. for baseline scenario, 20 p. for global scenario
- all fill out a custom-designed questionnaire
- independent variables: explanation type
- control variables: demographic variables, gaming experience, risk propensity, trust propensity, utilitarianism
- dependent variables: trust (capacity and moral) and XAI satisfaction

Results

- No significant differences were found between the baseline and global explanation scenario.
- Similarly high values for trust and XAI satisfaction



Discussion & Conclusion

- Global explanations as good as baseline explanations – trust and XAI satisfaction consistently high.
- No clear advantage of one type of explanation over the other – possible reasons include user sample, already sufficient baseline explanations, reliable robot’s behavior
- Future research – investigate global explanations across more diverse user populations and different scenario to examine its effectiveness