Explainable AI for Human Supervision over Firefighting Robots
How Do Textual and Visual Explanations Affect Human Supervision and Trust in the Robot?

Bogdan-Constantin Pietroianu

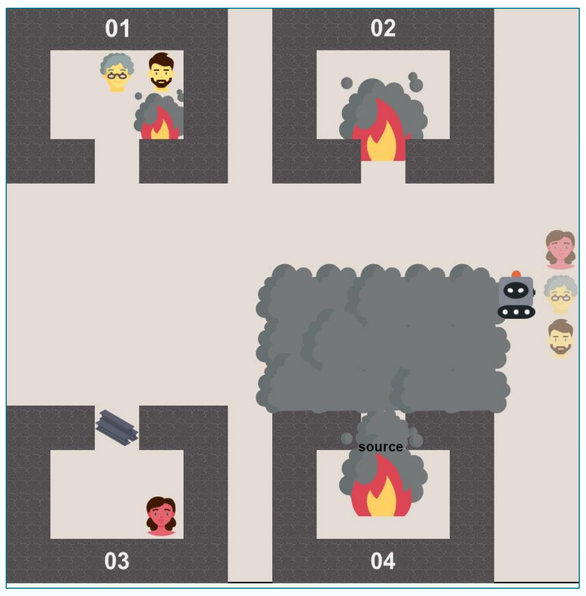
Delft University of Technology, b.c.pietroianu@student.tudelft.nl

1. Introduction

- Address the black-box design of AI models
- Presentation of the decision-making process in morally sensitive situations
- How does explanation type affect user trust
- Comparing visual and textual explanations

2. Task

- Search and rescue victims
- Extinguish fires
- 14 offices
- 11 victims
- Situational variables
- User intervention

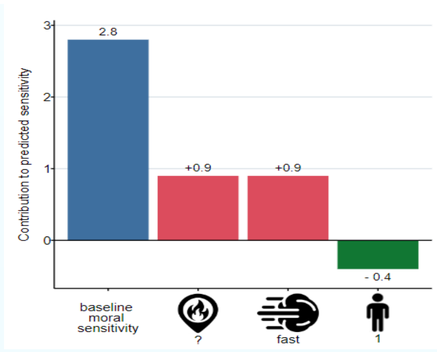


3. Method

- User study with 40 participants
- Measure demographics for result significance
- Questionnaire and experiment logs

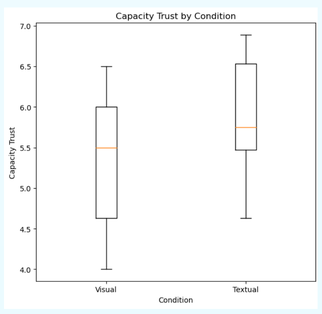
4. Explanations

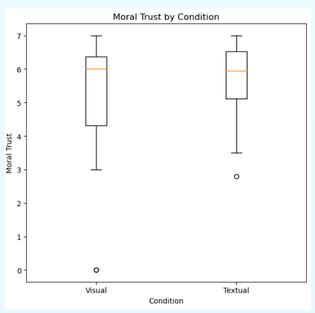
Brutus: I have found 👩, 👨, 👴 in the burning office 3. We should decide whether to first extinguish the fire or evacuate the victims. Please make this decision because the predicted moral sensitivity of this situation (**4.4**) is above my allocation threshold. This is how much each feature contributed to the predicted sensitivity:

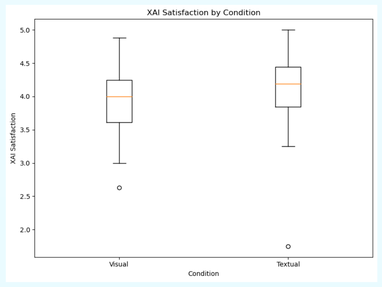


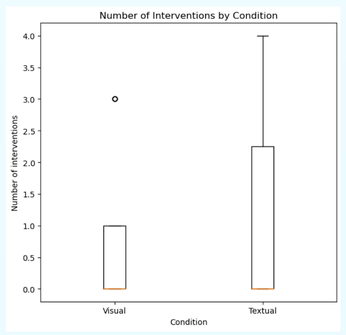
- The baseline moral sensitivity is **2.8**.
- The source of the fire is **unknown**, which **adds 0.9** to the baseline moral sensitivity.
- The speed at which the smoke is spreading is **fast**, which **adds 0.9** to the baseline moral sensitivity.
- The number of victims found is **1**, which **subtracts 0.4** from the baseline moral sensitivity.

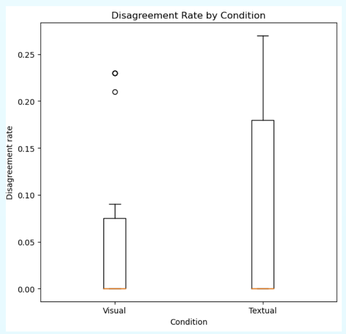
5. Results











6. Discussion and Conclusion

- Textual explanations yielded higher trust and involvement
- Analysis of demographics highlighted limitations in the generalizability of results
- The results highlight the importance of explanation modality in AI-human interactions

TU Delft | AI Initiative