Author:
Yi Wu
y.wu-37@student.tudelft.nl

Contrastive explanations for firefighting robots
User study to compare utility

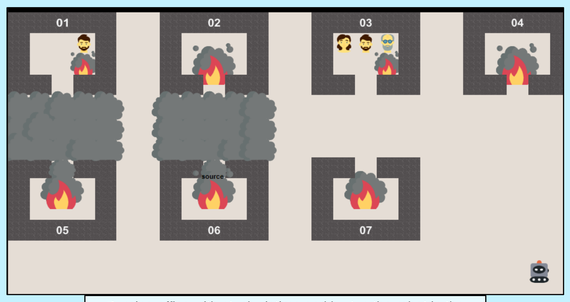
Responsible Professor:
Myrthe L. Tielman
Supervisor:
Ruben S. Verhagen

1. Background

Human-Agent Teamwork in Firefighting Operations

The integration of autonomous robots and human supervisors in firefighting scenarios enhances operational efficiency by automating routine tasks and reserving critical decision-making for human intervention. This study highlights the capabilities of Brutus, an advanced firefighting robot, within a simulated rescue mission framework.

- ❑ **Decision Allocation**: Brutus autonomously determines whether to make decisions itself or defer to the human supervisor.
- ❑ **Moral Sensitivity-Based Decisions**: Decisions are allocated based on a calculated moral sensitivity index, which considers various situational variables.
- ❑ **Variable allocations:** The variables and sensitivity calculations used in this study are derived from previous research conducted on the Brutus firefighting robot.



Burning office with stuck victims and brutus the robot in the bottom right corner.

2. Explainable Artificial Intelligence (XAI)

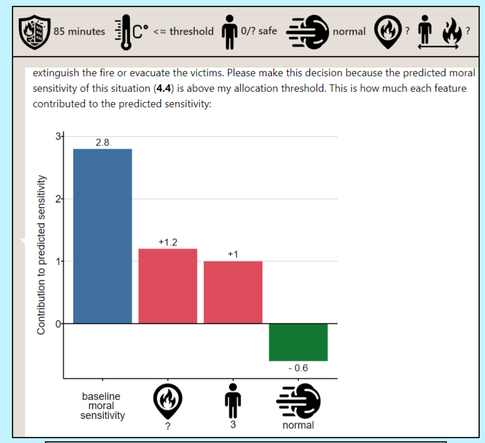
Communication and Decision-Making Process

The human supervisor and Brutus, the firefighting robot, communicate through a chatbox interface. This interface allows Brutus to provide real-time updates on its pathing, choices, and current actions.

Graphical Representation: When a decision needs to be allocated, a graph is displayed, illustrating the weight of each variable involved in the decision-making process.

Variable Combination: Depending on the situation, a combination of the following variables will be used:

* Time elapsed
* Temperature
* Distance between victim and fire
* Number of victims in the room
* Smoke spread
* Victim Locations



Message of the graph showing the weights of each variable.

3. Scenario, contrastive explanations

Studies have demonstrated that providing a contrastive view enhances user understanding of the explanations read. The current explanation outlines the robot's decision and variable allocations but lacks the reasoning behind these allocations. A contrastive explanation can be achieved by presenting the alternative allocations that would lead the robot to make the contrastive decision.

The contrastive decision can be identified when the calculated moral sensitivity surpasses the static moral threshold. By using a Breadth-First Search on the combination of variable allocations, we can determine the minimal changes needed to reach the contrastive decision.

This research aims to determine how such contrastive explanations, provided through alternative allocations, will influence human trust and supervision over the robot.

Given these change(s) I would allocate the decision to **myself**:
If the smoke spread was **normal** instead of **fast**.
If the fire location was **known** instead of **unkown**.
If we had **more time** than **44** min.

Contrastive explanation added to the existing graph message.

4. Methodology

Study Design: A between-subjects experiment was conducted with 40 participants.

Participants: The 40 participants were split into two groups of 20 (45% female, 55% male).
37 participants are aged 18-24, and
3 participants are aged 25-34.

Procedure:

Baseline Group: This group conducted the simulation with general explanation generation only.

Contrastive Explanation Group: This group conducted the simulation with general explanation generation plus added contrastive explanations.

Surveys: To measure subjective measures we use existing questions for:

- ❑ Moral trust
- ❑ Capacity trust
- ❑ XAI satisfaction

Lastly, we keep track of the disagreement rate measured during the experiment.

5. Results and conclusion

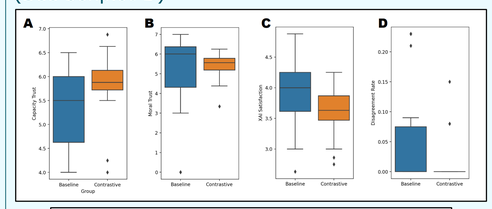
For trust and satisfaction, we expect higher outcomes for the contrastive group compared to the baseline group. For the disagreement rate, we do not assume a higher or lower rate for the baseline group. The independent samples t-test will be used if the assumptions are met; otherwise, the Wilcoxon test will be applied.

For capacity trust test showed that
there was a significant capacity trust difference between the baseline (M = 5.38, SD = 0.76)
and the contrastive (M = 5.82, SD = 0.67), W = 142, p = 0.029. (See boxplot A)

For moral trust the test showed no significant moral trust difference between the baseline (M = 5.05, SD = 2.036) and the contrastive (M = 5.39, SD = 0.66), W = 102, p = 0.55. (See boxplot B)

For XAI satisfaction the test showed no significant satisfaction difference between the baseline (M = 3.89, SD = 0.56) and the contrastive (M = 3.59, SD = 0.40), t(38) = 1.97, p = 0.97. (See boxplot C)

For the disagreement rate the test
showed no significant disagreement rate difference between the baseline (M = 0.06, SD = 0.090) and the contrastive (M = 0.016, SD = 0.04), W = 11.0, p = 0.092. (See boxplot D)



Boxplots of Capacity trust (A), Moral trust (B), XAI satisfaction (C) and Disagreement rate (D)

Conclusion

Results indicate that contrastive explanations significantly increased participants' capacity trust in the robot, though they did not significantly affect moral trust. Additionally, the results showed a lower satisfaction level with the explanations given by the robot. The disagreement rate between human decisions and robot actions was lower in the contrastive group, suggesting possible enhanced understanding and agreement with the robot's decisions.

These findings underscore the potential of contrastive explanations to enhance trust and collaboration in human-robot teams, paving the way for more effective integration of robots in critical operations. Future research should focus on larger sample sizes and explore the inclusion of contrastive decisions made by the robot alongside explanations to further validate these findings