<The influence of robot explanations on human-robot teamwork for firefighting: Adding contrastive explanations to feature attributions >

< **Yi Wu**[1]>
Supervisor(s): <Myrthe L. Tielman[1]**>, <Ruben S. Verhagen**[1]**>**
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: <Yi Wu>
Final project course: CSE3000 Research Project
Thesis committee: <Professor: Myrthe L. Tielman>, <Supervisor: Ruben S. Verhagen>, <Examiner: David Tax>

Abstract

The integration of robots in human-robot teams, particularly in high-stakes environments like firefighting, requires effective communication and decision-making to ensure safety and efficiency. This study explores the impact of adding contrastive explanations to feature attributions in robot explanations on human-robot teamwork during firefighting simulations. Contrastive explanations aim to improve human understanding by highlighting why a robot chose one decision over another using allocations of variables. The experiment involved 40 participants, divided into two groups, each interacting with either the baseline or contrastive version of the robot in the simulated environment. 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.

1 Introduction

The field of artificial intelligence (AI) has witnessed a remarkable surge in interest, increasing research and experimentation into the topic. One notable area of exploration is robot autonomy. While there are fully autonomous robots capable of completing simple routine tasks, such as those in industrial manufacturing, current technology does not support autonomous robots that can reliably perform complex high-level tasks and adapt to unforeseen situations. Therefore, integrating human-robot teams presents a viable and promising alternative.

Human-agent teamwork is seen as a better alternative to fully autonomous robots [1]. This approach allows robots to make and perform routine decisions while enabling human supervisors to oversee the robots' actions. When robots encounter situations requiring extra care or presenting high risks, they can defer to humans for the final decision.

This dynamic has led to research focused on replacing or enhancing tedious and dangerous tasks with AI. A prominent example is the deployment of firefighting robots to assist in hazardous environments. In these scenarios, human supervisors can oversee the decisions made by the firefighting robots, ensuring that potential mistakes are caught or prevented.

For effective human-robot teamwork, clear communication is essential to build trust [2]. This requirement aligns with the principles of explainable AI (XAI) [3], emphasizing the need for robots to provide understandable explanations of their decisions and actions, as well as to comprehend human feedback accurately.

This brings us to a study focused on this specific scenario [4]. In this study, the firefighting robot makes decisions and communicates the current circumstances and reasoning behind its choices to the human supervisor. While this research is thorough and well-executed, it has limitations regarding the generation of explainable AI (XAI). Research into explanation generation highlights several important factors, one of which is the "why" question

[5, 3]. Current methods often explain circumstances through feature attributions but lack the context to clarify why these features lead to specific decisions.

The concept of contrastive explanations addresses this gap by providing additional context that answers the "why" question [5, 3]. Studies have shown that contrastive explanations enhance understanding by elucidating what alternative decisions could have been made under different circumstances [6, 7, 8, 9]. By providing the alternative circumstance, or allocations of variables, under which the robot would have made a different decision, we can offer a contrastive explanation. This research aims to determine how such contrastive explanations, provided through alternative allocations, will influence human trust and supervision over the robot.

To answer this question we will be conducting a user study, but first, we will discuss the firefighter robot implementation in more detail and the related paper. Only then can we get into the details of the user study, its results, and the conclusions we can extract from the data.

2 Background

In this chapter, we will explore the background of our study, which builds upon an existing firefighting robot simulation. This foundation provides the context and basis for our current research.

2.1 Firefighting robot

As the application of AI expands across more technologies and tasks, effective communication becomes crucial, underscoring the importance of explainable artificial intelligence (XAI) [3, 5]. In addition to robust AI algorithms, the generation of clear explanations and visuals is essential for successful human-robot collaboration.

Our focus is on the firefighting robot described in [4]. This robot's algorithm and decision-making processes are informed by user studies and research on trust between human and robot agents [2]. The firefighting robot operates in a simulated environment where conditions are too hazardous for human firefighters, and its primary task is to rescue all civilians, including the injured, trapped in the burning building.

The robot communicates all its decisions and the steps taken to reach them. However, communication is not one-sided; humans can intervene and respond when a decision is allocated to them. We are using a Team Design Pattern (TDP) called Coactive Moral Decision Making (TDP3) [1]. Decisions can be made by either the human or the robot, with the human having the ability to intervene in any decision made by the robot.

2.2 Contrastive explanations

Something the human agent might question when working with the robot and reading its explanations is, why was this decision made with these allocations of variables? Questioning why X happened instead of Y in a situation is an important question that can be answered by providing a contrastive explanation. When confronted with a decision made by the

robot, the human agent might have expected a different choice. A contrasting view can aid in providing a better understanding for the human agent as to why the robot made such a decision. [5, 7, 6]

It is not the question of why did X happen for which contrastive explanations are needed, but the question of why did X happen instead of Y. This distinction is crucial, as the decision made by the robot, referred to as X, can be explained straightforwardly [7]. The reason decision X was made is due to the moral sensitivity not surpassing the moral threshold under the current circumstances. This is illustrated in the graph provided by the robot, detailing the weight of each feature variable (see Figure 1). However, when looking at just the graph, the human agent might wonder why these variable allocations resulted in this specific moral sensitivity calculation. Providing the alternative allocations that would lead to a different, contrastive decision offers the human agent the context needed to understand how these allocations influence the moral sensitivity, making it either better or worse.

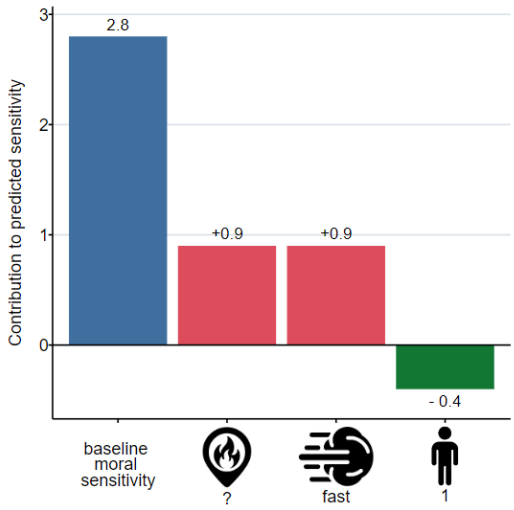

Figure 1: Graph sent by the robot containing the weight of each variable allocation towards the moral sensitivity.

In addition to answering the "why" question, a study by Kayo Yin and Graham Neubig concluded that *"Overall, contrastive explanations give a more intuitive and fine-grained interpretation of language models."* [10]. In their study, the contrastive explanation was used to determine why the language model predicted one token instead of another. This aligns with our study, as we aim to understand why the robot made one decision using variable allocations instead of another. This suggests a benefit to incorporating contrastive explanations into the base explanation for a better understanding by the human agent.

By providing this alternative view in the form of a contrastive explanation we will be able to answer this "why X instead of Y" question and give the human agent more understanding and hopefully trust for which they can make decisions for the robot or even disagree and intercept decisions made by the robot instead.

3 Method

With the background now established, we can delve into the specifics of the user experiment. This chapter will cover the experiment's design, participant details, the hardware and software used, the experiment environment, task, agent types, the generation of contrastive explanations, measurement methods, and the procedural steps undertaken.

3.1 Design

This user experiment employs a between-subjects design. Participants will be randomly assigned to one of two versions of the experiment, ensuring that each participant only experiences one version. The first version is based on this previous study by Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman [4] with some adjustments for the base explanations and will serve as the baseline. The second version introduces an additional contrastive explanation to the baseline, allowing us to compare the effectiveness of the enhanced explanation.

3.2 Participants

A total of 40 participants were recruited for this user experiment, all of whom are university students contacted through personal connections. The participants include 18 females and 22 males. The age distribution is as follows: 37 participants are aged 18-24, and 3 participants are aged 25-34. Regarding the highest level of education completed, 11 participants are high school graduates, 23 participants have some college credit but no degree, 1 participant holds an associate's degree, 3 participants hold a bachelor's degree, and 2 participants have obtained a master's degree. As for gaming experience, 6 participants reported having no experience, 7 participants had a little, 10 participants had a moderate amount, 7 participants had a considerable amount, and 10 participants had a lot of gaming experience.

Since the participants will be split into two groups, we will need to check for any significant differences. For gender, the ratio of men and women is equal, and therefore there are no differences (11 men and 9 women per group). As for age, education, and gaming experience, we used the Wilcoxon test. The results showed that there is no significant age difference between the groups for the baseline (Mean (M) = 1.05, Standard Deviation (SD) = 0.22) and the contrastive (M = 1.1, SD = 0.31), Wilcoxon statistic (W) = 0, p = 0.32. Neither is there a significant education difference between the groups for the baseline (M = 3.9, SD = 1.29) and the contrastive (M = 4.2, SD = 0.70), W = 40, p = 0.23. Nor is there a significant difference in gaming experience between the groups for the baseline (M = 3.15, SD = 1.60) and contrastive (M = 3.25, SD = 1.21), W = 68.5, p = 1.0.

Some variables we also take into account beforehand are Risk propensity, Trust propensity, and Utilitarianism. Assumptions for an independent samples t-test are not met for the Risk propensity, which is why we will use the Wilcoxon test. From the results, we can observe that there is no significant risk propensity difference between the groups for the baseline (M = 3.62, SD = 0.87) and the contrastive (M = 4.11, SD = 1.072), W = 55.5, p = 0.064. For the Trust Prospensity, the assumptions are met to perform an independent samples t-test, which shows us that there is no significant Trust propensity difference between the group for the baseline (M = 3.80, SD = 0.54) and the contrastive (M = 3.55, SD = 0.64), t-statistic(degrees of freedom: 38) = 1.36, p = 0.18. The assumptions are not met for

Utilitarianism and therefore we once again use the Wilcoxon test. The test shows us that there is no significant Utilitarianism difference between the group for the baseline (M = 2.89, SD = 0.58) and the contrastive (M = 2.84, SD = 0.51), W = 91.5, p = 0.89.

This shows that the two groups have no significant differences and are therefore equally distributed.

3.3 Hardware and Software

To conduct this user experiment, a laptop was used to run the software, which was coded in Python[1] and executed on the MATRX[2] platform. The laptop ran the software locally, initiating the simulation environment.

3.4 Environment

The simulation environment replicates a building office floor comprising 14 offices, some of which are on fire. One office is designated as the fire source, indicated by a small text label reading "source." A total of 11 injured civilians are dispersed throughout the offices, with mildly injured individuals marked in yellow and critically injured individuals marked in red (see Figure 2). The objective is to rescue all injured civilians, upon which the simulation will conclude.

A robot, the virtual agent named Brutus, is present on the burning office floor and serves as the primary rescuer with whom the human participant will collaborate. As the simulation progresses, obstacles may appear, and human firefighters can be dispatched to assist as needed.

On the top right of the interface, allocation variables are displayed and will continuously update as the situation progresses. Below the variables, the main communication takes place through a chatbox, with interactions facilitated by buttons located at the bottom (see Figure 3).

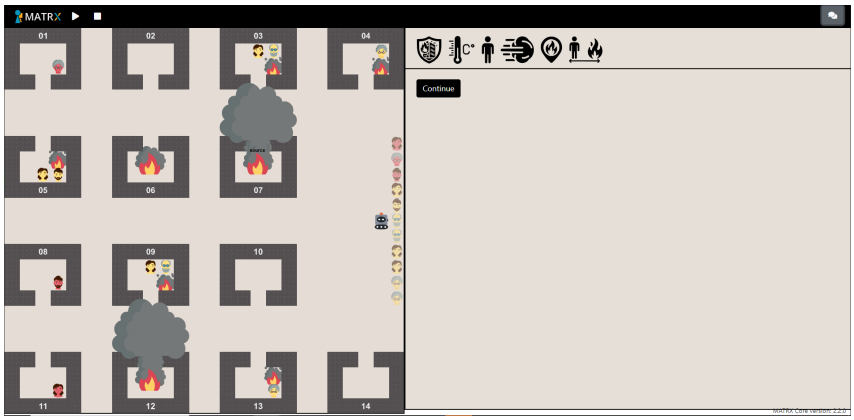

Figure 2: Main screen with the interface of the simulated world.

[1] https://www.python.org/
[2] https://matrx-software.com/

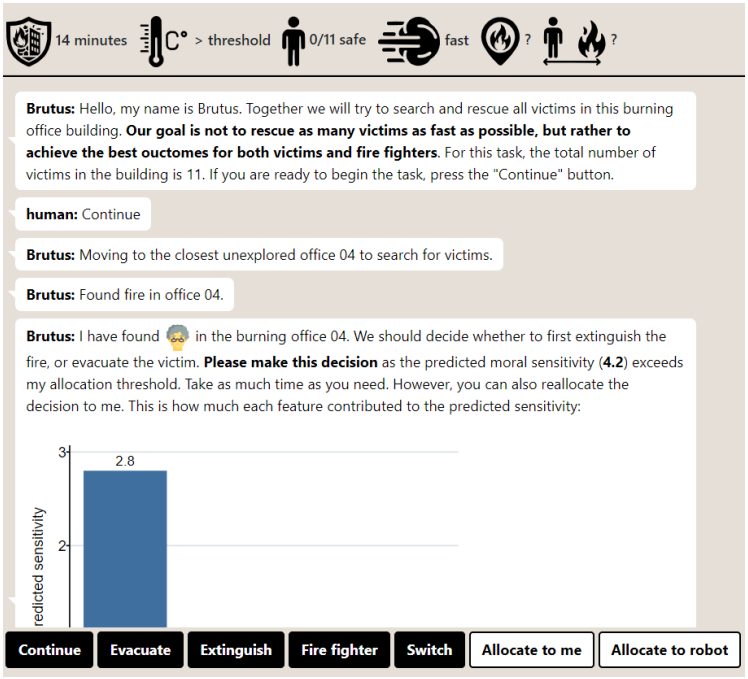

Figure 3: Right side containing the variables, chatbox, and interaction buttons.

3.5 Task

The primary task is for the human-robot team to collaborate in finding and rescuing all injured civilians trapped in the burning offices. The robot will prompt the human participant to make decisions in four different situations:

1. **Operational Tactic:**
   - **Offensive:** Prioritize finding and rescuing civilians without extinguishing found fires.
   - **Defensive:** Prioritize extinguishing fires over rescuing civilians.
2. **Room with Victim and Fire:**
   - Decide whether to rescue the victim first or extinguish the fire in the room first.
3. **Locating the Fire Source:**
   - If the source of the fire is not yet located, decide whether to continue the current strategy or send in firefighters to help locate the fire source.
4. **Rescuing Critically Injured Victims:**
   - Determine whether it is safe enough to send a firefighter to rescue a critically injured victim.

These decisions will guide the robot's actions and influence the overall strategy and safety of the rescue operation.

3.6 Agent Types

There are two agents present in the simulation: the autonomous virtual agent robot, Brutus, and the human supervising the robot. Brutus uses an implemented algorithm to explore rooms and plan routes to said rooms. When one of the four previously discussed decisions arises, the robot will first estimate the moral sensitivity. If the static moral threshold is not crossed, Brutus will make the decision autonomously; otherwise, it will defer the decision-making to the human agent. Additionally, the human agent has the ability to override Brutus's decisions and allocate the decision-making to themselves if they disagree with the robot's choices.

3.7 Contrastive Explanation Generation

As seen in Figure 1, the graph displays different allocations whose values change depending on the current situation. The allocations are represented by their current state, either through a label or a number. Without context, particularly for numerical values, it is difficult to ascertain whether these values are too high, too low, good, or bad.

The current situation provides a moral sensitivity value that determines whether the robot will make a decision autonomously or defer to the human. By using the same algorithm to calculate moral sensitivity, we can find the next best value at which the opposite decision would be made, thereby providing a contrastive perspective. Each allocation influences the moral sensitivity. Using a Breadth-First Search (BFS) over the combination of allocations, we can identify the next best combination of changed allocations that would lead to the contrastive choice.

To make the allocations simple for the human agent to understand, we will provide additional explanations using the new allocations to offer context. This result is illustrated in Figure 4.

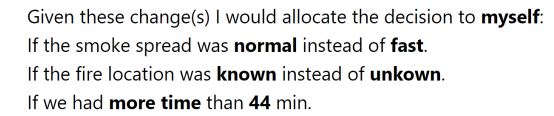

Figure 4: Changed allocated variables for a contrastive explanation.

3.8 Measures

As we aim to determine whether the provided extra context is helpful and its impact on human trust, we will be noting the subjective measures from our users. These subjective measures include the Risk Propensity Scale [11], the Propensity to Trust Scale [12], the Utilitarianism Scale [13], the Multi-Dimensional Measure of Trust (MDMT) [14], and the Explainable Satisfaction Scale [15].

The Risk Propensity Scale contains 7 questions, each scored from 1 to 9, with higher scores indicating a higher propensity for risk-taking. The Propensity to Trust Scale consists of 6 questions, scored from 1 to 5, with higher scores indicating a greater trust in technology. The Utilitarianism Scale comprises 9 questions, scaled from 1 to 5, with higher scores indicating a greater inclination towards deontology. The MDMT includes 16 questions divided into

two categories: 8 questions assess trust capacity and the other 8 assess trust morality. All 16 questions are scored from 1 to 7, with higher scores indicating greater trust in the robot's capacity and morality. Lastly, the Explainable Satisfaction Scale consists of 8 questions, scored from 1 to 5, with higher scores indicating higher satisfaction.

As for objective measures, we keep track of the disagreement rate observed during the experiment.

We will calculate the mean of the total score for each scale, allowing us to find any significant differences in trust and capabilities resulting from the experiments.

3.9 Procedure

The group of 40 participants was split into two groups: the first group conducted the baseline experiment, while the second group performed the experiment with the added contrastive explanation. The experiment began with surveys assessing their sense of risk, trust in technology, and utilitarianism. Following this, the respective experiment was performed and recorded. Afterward, participants filled out surveys based on their perception of the robot during the experiment, which were scored using the MDMT and Explainable Satisfaction Scale. The entire duration of the experiment was approximately 30 minutes.

4 Results

In this chapter, the observed values for trust, satisfaction, and disagreement rate are shared, along with the tests to check for any significant changes. 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.

4.1 Trust result

To observe the trust of the human agent towards Brutus the robot, we examined the capacity trust and moral trust survey results. For capacity trust, the assumptions for an independent samples t-test were not met, so the Wilcoxon test was used instead. The 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 Figure 5A)

Similarly, the assumptions for the independent samples t-test were not met for moral trust, so the Wilcoxon test was used again. 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 Figure 5B)

4.2 Satisfaction result

As for the satisfaction of the human agent towards Brutus the robot, we examined the XAI satisfaction survey results. The assumptions for the independent samples t-test were met, and thus the t-test was used to determine if there was any significant difference in the data.

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 Figure 5C)

4.3 Disagreement rate result

The disagreement rate is based on the times the human agent intervened in the decisions made by the robot, which we recorded for each experiment. The assumptions for the independent samples t-test were not met, so the Wilcoxon test was used instead. 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 Figure 5D)

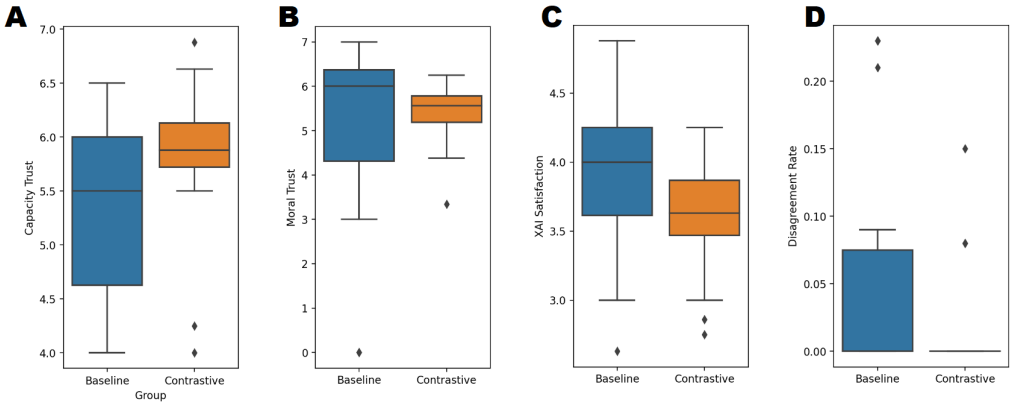

Figure 5: Boxplots of Capacity trust (A), Moral trust (B), XAI satisfaction (C) and Disagreement rate (D)

5 Responsible Research

5.1 Ethical issues

The participants were a crucial part of this study; therefore, great care was taken to ensure they were well-informed about all aspects of the study and their privacy was protected. The surveys were anonymized to safeguard participants' privacy, and participants were informed of this. Consent was obtained from participants at the start of the survey, and details of the study were provided. Participants were also informed that the data collected would be kept for possible future non-commercial research. Thus, at the start of the survey, all participants were informed and consented to the study, the usage of the data, and the anonymity measures in place. Additionally, to ensure the fairness of the research, we did not selectively choose our participants and retained all the results obtained. This approach ensures that the results we report are reflective of our observations.

Compared to many controversial topics surrounding AI, this study aims to enhance the support AI provides to humans. This promotes teamwork between human and robot agents, avoiding the contentious issue of full robot autonomy that most controversies focus on. This does raise the possibility that the interaction between humans and AI could become too convenient, leading to an over-reliance on this technology and creating excessive dependence.

One such example is the popular tool ChatGPT[3]. Apart from over-reliance, it is important to consider the potential for abuse by individuals with ill intentions, depending on who creates and utilizes this technology. Preventing this is a topic of discussion among scientists and philosophers alike, for which there is no definitive answer yet.

5.2 Experiment Reproducibility

The foundation of the current experiment is based on a previous implementation available in a public repository[4]. This ensures that anyone can access the project and follow the steps outlined in this paper to reproduce the experiment.

6 Discussion

6.1 Results Reflection

The results show a significant increase in capacity trust for the contrastive experiment, while there is no significant difference in moral trust. Satisfaction appears to be slightly lower for the contrastive experiment. The disagreement rate was tested using a two-tailed test, which showed no significant difference; however, the results in Figure 5D indicate a lower rate for the contrastive experiment compared to the baseline.

In general, the participants seem to trust the robot, as the results show a high average for both experiments, with the average being slightly higher for the contrastive experiment. This suggests that a contrastive explanation aids in understanding and increases the trust human agents have in the robot. From the standard deviation (SD) and the figure, we also observe a lower variance for the contrastive experiment. This suggests a more consistent trust level for contrastive explanations.

The general satisfaction of the participants seems to be lower for the contrastive explanation compared to the baseline. This could suggest dissatisfaction with the robot's decisions now that the human agent can understand why the current allocations result in those decisions instead of the other contrastive ones.

Although no significant difference in the disagreement rate was shown, this test was two-sided and chosen because the disagreement rate could have gone in either direction with the addition of contrastive explanations. The reason is that a contrastive explanation should lead to a better understanding, for which the human agents could either agree or disagree more with the robot. However, from Figure 5D, as well as the mean and SD, we can see a clear decrease in the disagreement rate for the contrastive experiment. This does show a higher agreement ratio for the human agents who participated in the contrastive experiment version, which could be attributed to a better understanding.

6.2 Limitations and Future Work

As we have mentioned the importance of participants, it is also necessary to consider that our participants are not related to the relevant job field for which this robot and simulation

[3] `https://chatgpt.com/`

[4] https://github.com/rsverhagen94/TUD-Research-Project-2024

are designed. Since none of our participants are firefighters, this influences the conclusions we can draw from this experiment, as seasoned firefighters would have better knowledge and understanding of the circumstances in the simulation. The reasons why participants might trust the robot more could be the same reasons some firefighters might distrust the robot.

Additionally, the participants vary in experience, age, and education, meaning there is no homogeneity in the data. Although, as shown before, there are no significant differences, it still indicates that participants have different ways of interacting and completing the simulation experiment.

There were some comments from participants who found the speed at which some messages were sent to be too slow, leading to dissatisfaction not related to the explanations themselves. Two participants also mentioned feeling somewhat dissatisfied with the difference in capitalization between human messages starting with "human: ..." and the robot's starting with "Brutus: ...".

Currently, only the alternative allocations for the contrastive explanation are provided. However, in the future, adding the robot's contrastive decision could enhance understanding. The potential drawbacks include information overload, as the human agent would need to process more information in an already time-dependent experiment. Studies have shown that reading attention and comprehension degrade when reading on-screen and under time pressure [16, 17]. Additionally, revealing the robot's decision could influence the decision the human agent might have otherwise made.

7 Conclusion

In this study, we investigated the influence of adding a contrastive explanation on the trust and satisfaction of human agents. The results show a significant increase in the trust level, indicating higher trust in the contrastive version. However, satisfaction seems to be lower, which could be due to various reasons. One possible reason is that, through better understanding via contrastive explanations, human agents may have disliked the robot's decisions more. Some participants also disliked the message speed of the robot, which is not related to the contrastive explanation but is an issue with the implementation of the robot itself, potentially causing a decrease in satisfaction. More research into satisfaction is needed as the current results are not clear enough to reach a definitive conclusion. Additionally, the results showed a decrease in the disagreement rate for the contrastive version. This could be attributed to a better understanding provided by the added contrastive explanations, but more research is needed to make a conclusive statement. Overall, the added perspective of a contrastive view does show a difference in how participants view and comprehend the explanations. However, the observed differences are not significant enough to make conclusive statements at this time, and further research with a larger sample size would help in identifying any possible significant differences. Future work could also expand on the contrastive explanations with added contrastive decisions made by the robot instead of just the contrastive allocations.

References

[1] J. van der Waa, J. van Diggelen, L. C. 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, Proceedings*, 2020, final published version.

[2] R. S. Verhagen, M. A. Neerincx, C. Parlar, M. Vogel, and M. L. Tielman, "Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance," in *Proceedings of the 2023 International Conference of Autonomous Agents and Multiagent Systems*, 2023, accepted author manuscript.

[3] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *IEEE*, 2020.

[4] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, *Meaningful human control and variable autonomy in human-robot teams for firefighting. Frontiers in Robotics and AI*, 2024.

[5] T. Miller, *Explanation in artificial intelligence: Insights from the social sciences*, 2018.

[6] J. Hoffmann and D. Magazzeni, "Explainable ai planning (xaip): Overview and the case of contrastive explanation (extended abstract)," in *Reasoning Web. Explainable Artificial Intelligence*, M. Krötzsch and D. Stepanova, Eds. Springer, Cham, 2019, vol. 11810, pp. 123–138. [Online]. Available: https://doi.org/10.1007/978-3-030-31423-1_9

[7] P. Lipton, "Contrastive explanation," *Royal Institute of Philosophy Supplement*, vol. 27, pp. 247–266, 1990.

[8] T. Miller, "Contrastive explanation: A structural-model approach," December 2020.

[9] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, *A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence*, 2021.

[10] K. Yin and G. Neubig, "Interpreting language models with contrastive explanations," *arXiv preprint arXiv:2202.10419*, February 2022, version 2, last revised 23 May 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2202.10419

[11] R. M. Meertens and R. Lion, "Measuring an individual's tendency to take risks: The risk propensity scale," *Journal of Applied Social Psychology*, vol. 38, no. 6, pp. 1506–1520, June 2008, first published: 21 May 2008. [Online]. Available: https://doi.org/10.1111/j.1559-1816.2008.00357.x

[12] S. M. Merritt *et al.*, "I trust it, but i don't know why: effects of implicit attitudes toward automation on trust in an automated system," *Human Factors*, June 2013. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/

[13] G. Kahane, J. A. C. Everett, B. D. Earp, L. Caviola, N. S. Faber, M. J. Crockett, and J. Savulescu, "Beyond sacrificial harm: A two-dimensional model of utilitarian psychology," *Psychological Review*, vol. 125, no. 2, pp. 131–164, March 2018, published online 21 Dec 2017. [Online]. Available: https://doi.org/10.1037/rev0000093

[14] D. Ullman and B. F. Malle, *MDMT: Multi-Dimensional Measure of Trust*, April 2019, version Date: 2019-04-01.

[15] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance," *Frontiers in Computer Science*, vol. 5, February 2023, published on 06 February 2023, Sec. Theoretical Computer Science. [Online]. Available: https://doi.org/10.3389/fcomp.2023.1096257

[16] R. Ackerman and T. Lauterman, "Taking reading comprehension exams on screen or on paper? a metacognitive analysis of learning texts under time pressure," *Computers in Human Behavior*, vol. 28, no. 5, pp. 1816–1828, September 2012. [Online]. Available: https://doi.org/10.1016/j.chb.2012.04.023

[17] P. Delgado and L. Salmerón, "The inattentive on-screen reading: Reading medium affects attention and reading comprehension under time pressure," *Learning and Instruction*, vol. 71, p. 101396, February 2021. [Online]. Available: https://doi.org/10.1016/j.learninstruc.2020.101396