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[1]
Supervisor(s): Dr. Myrthe L.Tielman[1]**, Ruben 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: Bogdan-Constantin Pietroianu
Final project course: CSE3000 Research Project
Thesis committee: Dr. Myrthe L.Tielman, Ruben Verhagen, Dr. David Tax

Abstract

As artificially intelligent agents become integrated into various sectors, they require an analysis of their capacity to make moral decisions and of the influence of human supervision on their performance. This study investigates the impact of textual feature explanations on human supervision and the trust in a semi-autonomous firefighting robot named Brutus, which operates in a morally complex environment. Grounded in the field of Explainable AI (XAI), which seeks to render AI decisions transparent, this research compares textual and visual explanations' effectiveness in conveying situational sensitivity during a simulated rescue operation. Through a detailed experimental setup using the MATRX software to simulate a burning office building, participants' trust and understanding were assessed based on their interaction with Brutus using either textual or visual explanations. This study contributes to the broader discourse on AI ethics and the optimization of human-agent teaming in high-stakes scenarios. The findings suggest that textual explanations can enhance human supervision and trust, fostering greater engagement and satisfaction compared to visual explanations.

1 Introduction

Artificially intelligent agents stand at the forefront of modern ethical discourse, particularly in their capacity to make moral decisions. As these agents increasingly integrate into various facets of our lives, from autonomous vehicles to virtual assistants, questions surrounding their moral agency and responsibility become paramount. These agents can operate with varying degrees of autonomy, either acting independently or relying on human supervision to make moral decisions [1]. Thus, it is crucial to focus on shaping the interaction between humans and machines to yield the best results. Research in this domain has explored the influence of interdependence in solving tasks [2], the effect of personalized explanations [3] and perceptions of autonomy [4]. This study aims to deepen our understanding of how communication with supervising humans can affect their perception of trust in tasks involving moral decisions.

Explainable Artificial Intelligence (XAI) has emerged as a critical field, driven by the necessity to make AI systems more transparent and comprehensible to users. XAI addresses the "black box" nature of many AI models by providing insights into how decisions are made, which is essential for debugging, improving AI models, and ensuring user trust [5]. Trust in AI systems is influenced by factors such as accuracy, perceived fairness, and the clarity of decision-making processes.

This study investigates the impact of textual feature explanations on human supervision and trust in a semi-autonomous firefighting robot named Brutus. Participants' trust and understanding were assessed based on their interaction with Brutus using either textual or visual explanations through a detailed experimental setup using the MATRX software to simulate a burning office building. The findings suggest that textual explanations can enhance human supervision and trust, fostering greater engagement and satisfaction compared to visual explanations. While participant characteristics varied, the insights gained contribute to the broader discourse on AI ethics and the optimization of human-agent teaming in high-stakes scenarios.

This paper will present the background of this study in Section 2 and the methods used by this research will be presented in Section 3. A consideration of responsible research can be found in Section 4. Section 5 will present a discussion about the findings of this research, the limitations of this study as well as possible future work. Finally, in section 7 the Conclusion will be presented.

2 Background

2.1 Explainable AI and Trust in AI Systems

Explainable Artificial Intelligence (XAI) has emerged as a crucial field of study within AI research, driven by the necessity to make AI systems more transparent and comprehensible to users. XAI aims to address the "black box" nature of many AI models, especially those relying on deep learning, by providing insights into how decisions are made. This transparency is crucial not only for debugging and improving AI models but also for ensuring user trust and fostering broader acceptance of AI technologies in various domains.

Trust in AI systems is a multi-faceted construct influenced by several factors, including the accuracy of the system, the perceived fairness, and the clarity of its decision-making processes. Among these, the ability of an AI system to explain its decisions in an understandable manner is paramount. Explanations can help users understand the rationale behind AI decisions, thereby increasing their confidence in the system and its outcomes [5]. This is particularly important in high-stakes environments where decisions carry significant moral and ethical implications, such as firefighting, healthcare, and criminal justice [6].

2.2 Visual Against Textual Explanations

One of the key debates within XAI is the comparative effectiveness of different types of explanations, particularly visual versus textual. Visual explanations typically include graphical representations such as feature importance graphs, or decision trees, which illustrate how different inputs influence the AI's decision. Textual explanations, on the other hand, provide written descriptions or summaries of the decision-making process [7].

Visual explanations are often lauded for their ability to convey complex information quickly and intuitively. For instance, a heatmap superimposed on an image can highlight which areas were most influential in the decision, making it easier for users to understand at a glance[8]. This can be particularly effective for users who are more visually oriented or when the information needs to be processed rapidly [9]. However, the effectiveness of visual explanations can be limited by the user's ability to interpret graphical data accurately, which may vary widely among different users.

Textual explanations, in contrast, can provide more detailed and precise information about the decision-making process. They are capable of delivering nuanced insights and reasoning that visual aids might oversimplify or omit. Textual explanations can be particularly valuable in scenarios requiring a thorough understanding of the decision rationale, allowing users to grasp the ethical and moral considerations involved in the AI's choices [10]. However, the clarity and usefulness of textual explanations are heavily dependent on the quality of the language used and the user's ability to comprehend complex textual information [11].

2.3 Comparative Analysis in Highly Moral Environments

In environments with high moral implications, the type of explanation provided by AI systems can significantly impact user trust. For example, in healthcare, where AI decisions can affect patient outcomes, both the transparency and comprehensibility of explanations are critical. Studies have shown that while visual explanations might enhance quick understanding and accessibility, textual explanations can offer more depth, which is crucial for making morally sensitive decisions [12]. Users in such environments might prefer detailed textual explanations to ensure that the AI's decisions align with ethical standards and legal requirements [13].

Moreover, the user's background and familiarity with the domain can influence their preference for visual or textual explanations. Experts might prefer detailed textual explanations that align with their need for thorough understanding, while non-experts might find visual explanations more approachable and easier to understand [14]. Therefore, the choice between visual and textual explanations should consider the target audience and the specific requirements of the high-stakes environment in question.

Both visual and textual explanations have their merits, their effectiveness in building user trust varies depending on the context and the user's needs. In highly moral environments, the depth and clarity provided by textual explanations may offer a significant advantage in ensuring that users feel confident in the AI's decisions and their alignment with ethical standards.

3 Methods

3.1 Design

An experiment was conducted to evaluate how different types of explanations will affect the perception and trust of users over a firefighting robot. It is split between two parts. The two alternatives covered by this research were visual and textual representations of the perceived situational sensitivity encountered in situations during a rescue operation from a burning building.

3.2 Participants

In this study, we investigated various demographic characteristics of participants under the visual and textual conditions. Understanding these demographics is crucial for interpreting the results and ensuring the generalizability of the findings. The following demographics were analyzed: gender, age, education, gaming experience, risk propensity [15], trust propensity [16], and utilitarianism [17]. The total sample analyzed for this experiment was n = 40, with 20 participants for each alternative.

In this study, gender was categorized into three groups: Female, Male, and Other/Prefer not to say. All of the participants identified as either female or male. The visual condition had 9 female participants and 11 male participants, while the textual condition had an identical distribution with 9 female participants and 11 male participants. A Chi-square test for homogeneity was conducted to compare the gender distributions between the two conditions. The test results showed no significant difference in gender distribution

($\chi^2(1, N = 40) = 0.0$, $p = 1.0$).

Age was categorized into seven groups: under 24, 25-34, 35-44, 45-54, 55-64, 65+ and Prefer not to say. The majority of participants were under 24 years old. In the visual condition, 19 participants were under 24, and 1 participant was in the 25-34 age group. In the textual condition, there were 16 participants under 24, 3 participants in the 25-34 age group, and 1 participant in the 45-54 age group. The Wilcoxon rank-sum test was conducted to compare the age distributions between the two conditions. The test results showed no significant difference in age distribution ($W = -0.83, p = 0.41$), indicating that the age distribution was similar between the two conditions.

Education was categorized into nine levels: no formal education, attending high school, high school diploma, attending college, associate degree, bachelor's degree, master's degree, PhD, and prefer not to say. In the visual condition, the distribution of education levels included 10 participants with a high school diploma, 7 attending college, 1 with a bachelor's degree, and 2 with a master's degree. In the textual condition, there were 2 participants with a high school diploma, 4 attending college, 13 with a bachelor's degree, and 1 with a master's degree. The Wilcoxon rank-sum test showed a significant difference in education levels between the two conditions ($W = -2.99, p = 0.003$), indicating that the educational backgrounds of participants in the two conditions varied significantly.

Gaming experience was categorized into five levels: none, a little, moderate, considerable, and a lot. The visual condition had 5 participants with no gaming experience, 2 with a little, 4 with a moderate amount, 3 with a considerable amount and 6 with a lot of gaming experience. The textual condition had 2 participants with no gaming experience, 2 with a little, 5 with a moderate amount, 5 with a considerable amount, and 6 with a lot of gaming experience. The Wilcoxon rank-sum test showed no significant difference in gaming experience between the two conditions ($W = -0.72, p = 0.47$).

Risk propensity was measured on a scale from 1 to 7, with higher scores indicating a greater propensity for taking risks. For the participants of the visual alternative the mean risk propensity score of 3.62 with a standard deviation of 0.84, while the participants of the textual alternative had a mean score of 4.11 with a standard deviation of 1.23. The Wilcoxon rank-sum test showed no significant difference in trust propensity between the two conditions ($W = -1.30, p = 0.19$).

Trust propensity was also measured on a continuous scale. The mean trust propensity for the visual condition was 3.80 with a standard deviation of 0.53, while for the textual condition, the mean was 4.08 with a standard deviation of 0.69. The Wilcoxon rank-sum test showed significant difference in trust propensity between the two conditions ($W = -1.96, p = 0.0498$).

Utilitarianism was measured on a scale from 1 to 5. The mean utilitarianism score for the visual condition was 2.89 with a standard deviation of 0.57, while for the textual condition, the mean was 3.13 with a standard deviation of 0.62. The Wilcoxon rank-sum test indicated no significant difference in utilitarianism between the two conditions ($W = -1.50, p = 0.25$).

In summary, while most demographic variables did not show significant differences between conditions, the education level and trust propensity did. These findings suggest that participants' educational backgrounds varied significantly between the two conditions, which could potentially influence the outcomes of the study. This will be further discussed in the limitations subsection.

3.3 Tools

During this study, a personal computer was used to log questionnaire responses and conduct the simulated rescue operation. The MATRX (huMan-Agent Teaming; Rapid eXperimentation) software was used to simulate a two dimensional environment in which the experiment was conducted.

3.4 Task

The MATRX environment used simulates a burning office building. A global overview of it can be seen in Figure 1. This is the "God" view where all the elements of the environment are visible at all times. The participants will be using a normal version in which victims and fires only appear on the map after the robot searched the office where they are located.

There are fourteen offices that need to be searched for finding and rescuing victims as well as extinguishing fires. There are two types of victims present: mildly injured (yellow) and critically injured (red). Mildly injured victims can be saved by the robot, which is named Brutus, individually while the critically injured victims require the intervention of a firefighter. There are two types of fire, small contained ones and a single source. The experiment monitors and accounts for several parameters relevant for the rescue operation. These variables are: the estimated remaining time until the building collapses, the current temperature compared to the safety threshold, the number of victims present, the speed at which the smoke spreads, whether or not the fire source has been found and the current distance of Brutus from the fire source. These elements are displayed at all times during the experiment and are relevant for the safety of firefighters and determine whether or not they can safely enter the building. During the rescue operations entrances might get blocked by falling debris which will need to be removed to continue the rescue operations.

Brutus can follow one of two strategies: offensive and defensive. In the offensive approach the priority is rescuing victims and extinguishing only fires that may impede this operation while the defensive one focuses on first extinguishing any fires and only then (re)start the rescue operation. The strategies can be alternated during the experiment.

3.5 Procedure

The experiment begins with a consent form which has to be agreed to by the participant in order for the study to continue. Afterwards the user has to respond to a set of questions about their demographics as presented in Section 3.2. The user is first introduced to the experiment by a tutorial to familiarise with the environment and how interactions with Brutus work. During this stage no model for the calculated situational sensitivity is shown to the user since the same tutorial is used during both the visual representation version and the textual one.

The main task is then started. The same version of environment is used for both alternatives. The visual model is presented in the form of a bar-chart as can be seen in Figure 2. The alternate textual explanation was constructed in a manner that provides the same contextual information. This decision was taken in order to ensure a fair comparison between the two alternatives as to only compare the method of presenting the information, not information itself. The textual explanation can be seen in Figure 3 and depicts the parameters for an identical situation as the one in Figure 2.

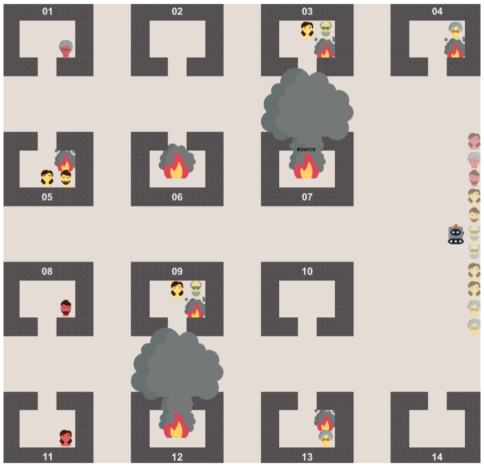


Figure 1: Overview of the rescue environment in the "God" view. All experiment elements are visible. This view was not shown to the user.

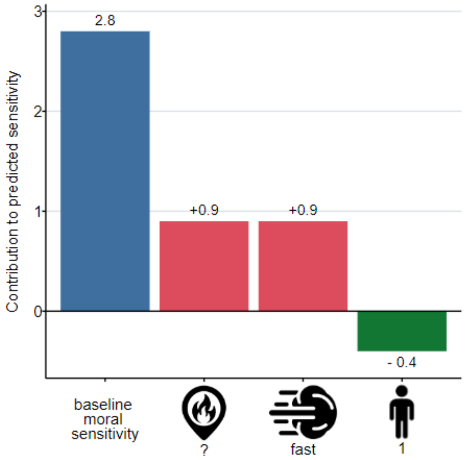
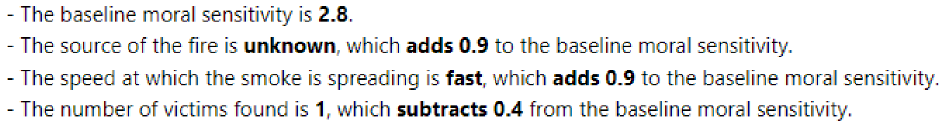


Figure 2: Visual way of explaining the user how the robot calculated the sensitivity of a situation.



- 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.

Figure 3: Textual way of explaining the user how the robot calculated the sensitivity of a situation.

3.6 Measures

A collection of subjective and objective measures was used to analyze the findings of this user study. The subjective matters were measured through the questionnaire after the completion of the experiment while the objective measures were logged by the software used during the experiment.

The subjective measures were capacity and moral trust[18] as well as XAI satisfaction[19]. The two trust measures were quantified with a set of eight questions each, measured on a scale from 0 to 7. There was also a 'Does not fit' option available for the respondent to select. The final value is represented by the average of all responses that are not marked as not fitting. The XAI satisfaction was measured as the average of a set of eight questions on a scale from 1 to 5.

The objective measures logged were completeness (the proportion of saved victims from all the victims), total allocations (number of total decisions made), human allocations (the number of times the participant was asked to make a decision), total interventions (the number of times a participant intervened in the robot's allocation of who will make a certain decision) and the disagreement rate (total interventions divided by total allocations).

4 Responsible Research

The utmost consideration was given to stringent ethical and responsible research practices to ensure the integrity, privacy, and transparency of this study.

Data anonymity was a key priority throughout the experiment. Identifiable information was not collected, ensuring that individual users could not be traced back to the data points. This approach aligns with the ethical guidelines outlined in the Netherlands Code of Conduct for Research Integrity, which emphasizes principles such as honesty, transparency, and responsibility.

Furthermore, to enhance the transparency and reproducibility of this research, the experimental repository is publicly available for independent evaluation[1]. This openness allows other researchers and reviewers to scrutinize our methodology, reproduce the experiment, and validate the findings. This practice supports the principle of transparency facilitating open and accountable research practices.

It is important to note that all data collected was fully used for the experiment. No data was discarded or excluded, ensuring a comprehensive analysis of the collected information. This inclusive approach strengthens the reliability and validity of the research outcomes, providing a complete picture of the experimental results.

Additionally, users had no prior knowledge about the experiment or how environmental variables were placed. This lack of prior information ensured that user behavior was natural and unbiased, thereby maintaining the authenticity of the data collected. The experiment's design aimed to observe genuine reactions and interactions within the given environment, free from any preconceived notions or expectations, aligning with the principles of independence and scrupulousness.

By following these guidelines and principles, this research not only adheres to national standards for research integrity but also contributes to the credibility and reliability of scientific findings.

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

5 Results

5.1 Trust Perception

For the dependent variables Capacity Trust and Moral Trust, data was analyzed using descriptive statistics and box plots to compare the effects of two conditions: Visual and Textual.

In the Visual condition, the mean Capacity Trust was 5.37, with a median of 5.50, a maximum of 6.50, a minimum of 4.00, and a standard deviation of 0.74. The mean Moral Trust in this condition was 5.05, with a median of 6.00, a maximum of 7.00, a minimum of 0.00, and a standard deviation of 1.98.

In the Textual condition, the mean Capacity Trust was 5.89, with a median of 5.75, a maximum of 6.89, a minimum of 4.63, and a standard deviation of 0.64. The mean Moral Trust was 5.73, with a median of 5.94, a maximum of 7.00, a minimum of 2.80, and a standard deviation of 1.13.

These statistics indicate that both Capacity Trust and Moral Trust were higher on average in the Textual condition compared to the Visual condition. Additionally, the greater variability in Moral Trust ratings in the Visual condition suggests that participants' perceptions of moral trustworthiness were more diverse when influenced by visual explanations.

The box plots in Figure 4 illustrate the distribution of Capacity Trust and Moral Trust scores across the two conditions. For Capacity Trust, the Textual condition showed a higher median score and a tighter interquartile range, indicating more consistent ratings. For Moral Trust, the Textual condition also had a higher median score, but the Visual condition had a wider range and interquartile spread, highlighting more variability in ratings.

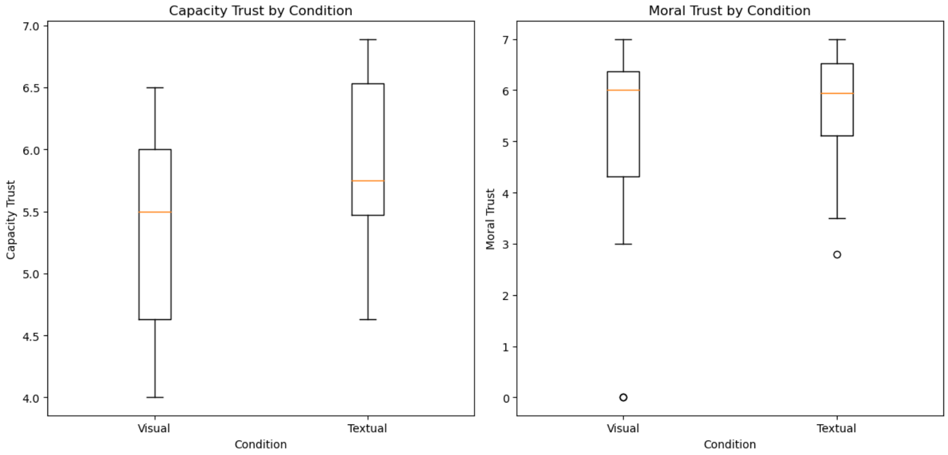


Figure 4: The distribution of the values reported for capacity and moral trust between the visual and textual alternatives.

5.2 Explanation Satisfaction

The mean XAI Satisfaction in the Visual condition was 3.89, with a median of 4.00, a maximum of 4.88, a minimum of 2.63, and a standard deviation of 0.54. In the Textual condition, the mean XAI Satisfaction was 4.08, with a median of 4.19, a maximum of 5.00, a minimum of 1.75, and a standard deviation of 0.73.

These statistics suggest that participants generally rated their satisfaction with the explanations higher in the Textual condition compared to the Visual condition. The higher variability in satisfaction ratings in the Textual condition suggests more diverse perceptions among participants regarding textual explanations.

The box plot in Figure 5 provides a visual representation of XAI Satisfaction scores. The median score was higher in the Textual condition, with a wider interquartile range indicating more variability in ratings.

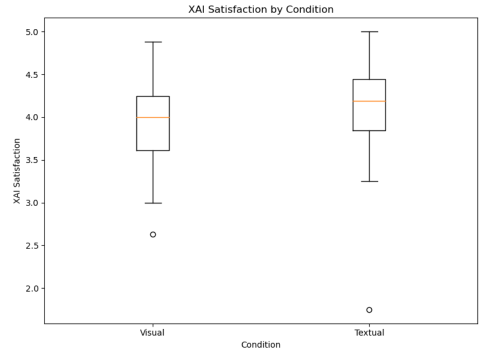


Figure 5: The distribution of the values reported for XAI satisfaction between the visual and textual alternatives.

5.3 Number of Interventions and Disagreement Rate

The mean number of interventions in the Visual condition was 0.80, with a median of 0, a maximum of 3, a minimum of 0, and a standard deviation of 1.17. The disagreement rate in the Visual condition had a mean of 0.06, a median of 0.00, a maximum of 0.23, a minimum of 0.00, and a standard deviation of 0.09.

In the Textual condition, the mean number of interventions was 1.10, with a median of 0, a maximum of 4, a minimum of 0, and a standard deviation of 1.37. The disagreement rate had a mean of 0.08, a median of 0.00, a maximum of 0.27, a minimum of 0.00, and a standard deviation of 0.10.

These results indicate that participants in the Textual condition had more interventions and a higher disagreement rate compared to those in the Visual condition. The higher variability in these metrics in the Textual condition suggests that participants were more active in intervening in the decision-making process.

The box plots in Figure 6 illustrate the distribution of the Number of Interventions and Disagreement Rate across the two conditions. For the Number of Interventions, the range was wider in the Textual condition, indicating more variability. Similarly, for the Disagreement Rate, the Textual condition had a wider interquartile range and overall spread, indicating greater variability in the disagreement rates.

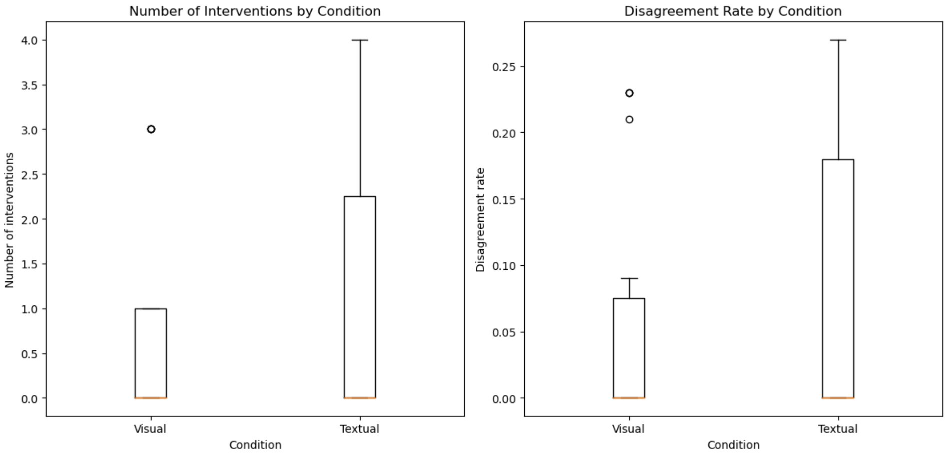


Figure 6: The distribution of the values reported for number of disagreements and disagreement rate between the visual and textual alternatives.

6 Discussion

6.1 Result Analysis

The findings from this study provide evidence that textual explanations enhance participants' trust in AI systems, especially in scenarios that involve morally complex decisions. The consistently higher scores for Capacity Trust and Moral Trust in the Textual condition suggest that participants found the textual explanations more reliable and easier to understand. This indicates that textual explanations might be more effective in conveying the AI's decision-making process, which is crucial in building trust.

Moreover, the narrower interquartile ranges for Capacity Trust in the Textual condition highlight that participants' trust ratings were not only higher but also more consistent. This consistency implies a general consensus among participants regarding the effectiveness of textual explanations in fostering trust. On the other hand, the wider variability in Moral Trust ratings for the Visual condition suggests that visual explanations might not be as universally effective, potentially leading to mixed perceptions among users.

The higher XAI Satisfaction scores in the Textual condition further emphasize the benefits of textual explanations. Participants felt more satisfied with the explanations provided in text form, which likely contributed to their higher trust ratings. This satisfaction could be due to the precise nature of textual explanations, which allow users to grasp the nuances of the AI's decision-making process better than visual explanations.

Interestingly, the higher number of interventions and disagreement rates in the Textual condition might initially seem like a drawback. However, this increased engagement can be interpreted positively. It indicates that participants were more actively involved in the decision-making process when provided with textual explanations. This active involvement is crucial in high-stakes environments, as it ensures that users are critically evaluating the AI's decisions rather than passively accepting them. Higher engagement levels can lead to better oversight and more informed decision-making, ultimately improving the overall performance and safety of human-agent teams. Despite the higher trust propensity in the Textual condition, the increased number of interventions suggests that participants were more vigilant and willing to question the AI's decisions, which can be seen as a sign of healthy skepticism and proactive behavior in ensuring the best outcomes.

These results highlight the importance of explanation modality in AI-human interactions. Textual explanations not only foster higher trust and satisfaction but also encourage more active engagement from users. This is particularly beneficial in morally complex environments, where understanding the rationale behind AI decisions is critical. Future research should explore the underlying reasons for these differences and investigate whether certain types of textual explanations are more effective than others.

It is important to note that the significant difference in education levels and trust propensity between the two conditions could have influenced the outcomes of the study. Participants in the Textual condition had a higher educational background and were more prone to trust technology on average, which might have contributed to their better understanding and higher trust in the AI explanations provided. These disparities should be considered when interpreting the results.

6.2 Limitations

Several limitations in this study should be considered when interpreting the results. The homogeneity of the participant group may affect the generalizability of the findings. Since the study involved a relatively uniform group of individuals, extending these results to a more diverse population may require further investigation.

Significant differences were noted in the educational backgrounds and trust propensities of participants in the visual versus textual conditions. Further regression analysis to discover the effect of these factors on the results was considered but could not be completed due to the limited time available. Participants in the textual condition had higher educational levels and a greater propensity to trust technology, potentially contributing to their better understanding and higher trust in the AI explanations provided. This disparity might influence the generalizability of the results.

The way samples were collected also posed a limitation. The observations for the visual alternative were collected before those for the textual one, as they were intended for additional use in other studies. This sequence limited the ability to distribute participants more evenly between the alternatives.

6.3 Future Work

The tools used in the experiment are extremely versatile and open up the path for further research in the domain of Explainable AI. The current experiment was designed with full completeness in mind in order to not affect the trust perception of users. New research can investigate the effect to which the completeness of a task influence the trust perception of participants. Additionally, the effect of the amount of information presented over user trust can also be investigated in future research.

7 Conclusion

This study indicates that textual explanations can enhance human supervision and trust in a semi-autonomous firefighting robot operating in morally complex environments. By comparing textual and visual explanations, the research found that participants in the textual condition reported a higher trust perception, along with greater satisfaction with the explanations provided. Additionally, textual explanations led to increased participant engagement, as evidenced by higher numbers of interventions and disagreement rates, suggesting more active involvement in the decision-making process.

These findings highlight the importance of explanation modality in AI-human interactions, particularly in high-stakes scenarios where understanding the rationale behind AI decisions is critical. Textual explanations appear to foster greater trust and satisfaction and encourage users to critically evaluate AI decisions, which is essential for ensuring the best outcomes in morally sensitive environments.

However, the study acknowledges certain limitations, such as the homogeneity of the participant group, differences in educational backgrounds and trust propensities between the visual and textual conditions. These factors should be considered when interpreting the findings.

This research contributes to the broader discourse on AI ethics and the optimization of human-agent teaming by providing insights into the effectiveness of textual explanations in building human trust and enhancing supervision in morally complex scenarios. As AI systems continue to evolve and integrate into various sectors, understanding how to communicate AI decision-making processes effectively will be crucial for fostering trust and ensuring ethical outcomes. Future research should continue to explore how different types of explanations can further improve human-AI collaboration.

References

[1] J. van der Waa, J. van Diggelen, L. Cavalcante 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, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22.* Springer International Publishing, 2020, pp. 203–220.

[2] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, "The influence of interdependence and a transparent or explainable communication style on human-robot teamwork," *Frontiers in Robotics and AI*, vol. 9, p. 993997, 2022.

[3] 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 on Autonomous Agents and Multiagent Systems*, 2023, pp. 2316–2318.

[4] J. van der Waa, S. Verdult, K. van den Bosch, J. van Diggelen, T. Haije, B. van der Stigchel, and I. Cocu, "Moral decision making in human-agent teams: Human control and the role of explanations," *Frontiers in Robotics and AI*, vol. 8, p. 640647, 2021.

[5] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[7] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[8] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[10] M. Hind, D. Wei, M. Campbell, N. C. Codella, A. Dhurandhar, A. Mojsilović, K. Natesan Ramamurthy, and K. R. Varshney, "Ted: Teaching ai to explain its decisions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 123–129.

[11] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–52.

[12] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 149–159.

[13] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[14] D. Gunning and D. Aha, "Darpaâs explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[15] R. M. Meertens and R. Lion, "Measuring an individual's tendency to take risks: the risk propensity scale 1," *Journal of applied social psychology*, vol. 38, no. 6, pp. 1506–1520, 2008.

[16] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, “I trust it, but i donât know why: Effects of implicit attitudes toward automation on trust in an automated system,” *Human factors*, vol. 55, no. 3, pp. 520–534, 2013.

[17] G. Kahane, J. A. 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, p. 131, 2018.

[18] S. Tolmeijer, M. Christen, S. Kandul, M. Kneer, and A. Bernstein, “Capable but amoral? comparing ai and human expert collaboration in ethical decision making,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–17.

[19] 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, p. 1096257, 2023.