CSE3000 Research Project

How do adaptive explanations that become more abstract over time influence human supervision over and trust in the robot?

Elena Ibanez[1]
Supervisor(s): 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: Elena Ibanez
Final project course: CSE3000 Research Project
Thesis committee: Myrthe L. Tielman, Ruben Verhagen, David Tax

Abstract

As human-agent collaboration grows increasingly prevalent, it is crucial to understand and enhance the interaction between humans and AI systems. Explainable AI is fundamental to this interaction, which involves agents conveying essential information to humans for decision-making. This paper investigates how adaptive explanations affect human supervision and trust in robotic systems. The study included 40 participants and compared baseline (non-adaptive) explanations with adaptive explanations. The results showed no significant difference between the two types of explanations; making explanations more abstract did not necessarily improve human supervision or increase trust in robots.

1 Introduction

The study of Human-Agent interaction is a rapidly growing field that offers numerous opportunities for creating agents that behave socially and communicate efficiently with humans [20]. Additionally, agents are becoming increasingly autonomous and intelligent, enabling them to perform tasks with minimal human supervision. However, humans should handle some sensitive circumstances exclusively, significantly when moral decisions in risky situations could affect people's safety [24]. In such cases, humans should intervene and assume responsibility for sensitive decisions while allowing the agent to handle decisions that are considered very safe independently [25].

These agents' rising complexity and capacities raise essential questions about their integration into environments where human safety and ethical considerations are important [21]. As agents become more embedded in dynamic environments, understanding their decision-making processes and ensuring they align with human values and ethical standards becomes essential. This intersection of advanced technology and ethical responsibility emphasizes the need for clear and effective communication between humans and agents, especially in high-stakes scenarios [2].

This research focuses on Human-Agent interaction within a firefighting context, particularly where moral decisions are involved. Depending on its predicted moral sensitivity, the agent must decide whether to allocate decision-making to itself or the human supervisor. For the human-agent team to succeed in this high-stakes environment, the clarity and personalization of agent explanations are crucial [3]. Furthermore, transparent and personalized explanations ensure that human supervisors understand the rationale behind the agent's actions and decisions, essential for maintaining trust and effective collaboration [13]. When human supervisors can comprehend the agent's decision-making process, especially in morally sensitive situations, they are better equipped to intervene appropriately and make informed decisions [3]. Hence, this paper aims to answer the question: "*How do adaptive explanations that gradually become more abstract affect human supervision and trust in the robot?*". To explore this further, the research addresses several subquestions: how should designing these explanations be approached, how can these explanations be generated and implemented in the agent, and how do these explanations affect the dependent variables, such as human supervision and trust?

2 Background research

2.1 Human-agent teamwork

Developments in AI make it possible for agents to work together with humans in a human-agent team (HAT) to accomplish shared objectives. However, delicate decision-making is usually considered a human competency [27]. As a result, when a human-agent team is involved in moral decision-making, the human takes responsibility for ensuring that ethical principles are respected and maintaining accountability if the team fails. For humans to collaborate with other agents, they must have significant control over them. Regardless, research has shown that human-agent interaction is only effective when a team environment values and supports the growth of human responsibility for the agent's actions and the group's choices [25]. In designing the team system, it is important to include explanations that match the team roles and the human cognitive state in how agents' actions and decisions are communicated and managed.

Furthermore, agents can perform various functions in human-agent interactions, including helping individuals complete tasks in collaborative settings and working independently to complete tasks [19]. For example, in environments such as firefighting, agents can assist human firefighters by entering dangerous areas to gather data, identify hazards, and perform initial safety measures, thereby reducing risk to human life, specifically the firefighters. Besides, agents can deliver supplies in natural disaster relief scenarios and assist in search and rescue operations, offering support where human presence is limited or dangerous [19]. Similarly, human-agent collaboration is critical in these settings, as it emphasizes designing agents as teammates rather than tools, highlighting the necessity of understanding behaviors that foster successful collaboration [18]. This approach ensures that agents enhance human-agent teams' efficiency, safety, and effectiveness across various applications.

Additionally, trust is a significant requirement for human-agent teams to be successful [11]. Trust in human-agent teams is developed through a history of successful interactions [28]. Developers of agent systems need to account for these social elements to ensure effective and fair teamwork. By addressing these factors, teams can function more effectively, leveraging the strengths of both humans and agents. Moreover, effective human-agent collaboration requires understanding the interdependencies in joint activities [11]. Ensuring all team members, including agents, know these interdependencies helps maintain transparency and fairness in task execution and decision-making.

Besides trust and understanding interdependencies, the configuration of human-agent teams plays a crucial role in optimizing their performance. One practical approach to configuring these teams is through dynamic task allocation. Dynamic task allocation involves continuously assessing and redistributing tasks among team members, including humans and agents, based on their current capabilities, workload, and situational demands [6]. For instance, capability-based task allocation involves assigning tasks based on humans' and robots' specific strengths and weaknesses [22]. This method has been validated in practical settings, demonstrating its effectiveness in improving work quality and efficiency by matching tasks to the most suitable resource.

Moreover, a team's efficiency can be significantly improved by dynamically reallocating tasks based on current conditions and cognitive load [7]. This approach allows for adaptive automation, where tasks are assigned to agents or humans according to their current capabilities and workload. Such flexibility is particularly beneficial in high-demand environments like disaster response, where the situation can change rapidly. In dynamic task allocation within human-agent teams, effective agent explanations are also recognized as a

crucial element for optimal performance and trust [14]. Explanations can also effectively improve human teammates' understanding of task allocation decisions by their agent teammates. Dynamic task allocation ensures the team can respond to these changes effectively, maintaining high performance and reducing the cognitive burden on human team members [7].

2.2 Explainable AI

Explaining decisions is integral to human communication, understanding, and learning [8]. XAI refers to methods and techniques in artificial intelligence that make the outputs of AI systems understandable to humans. Research highlights the importance of stakeholders selecting appropriate XAI approaches and tools based on their specific needs and the characteristics of the AI applications they are developing [5]. XAI is crucial when understanding the rationale behind an AI decision, such as in medical diagnosis or any industry that significantly influences human lives, like firefighting. If the behavior of a respective agent is not explained, the human may reach an explanation that does not necessarily reflect the AI's actual internal state. This can lead to self-deception and lower the quality of the interaction [1]. This could sometimes lead to dangerous situations, putting the human's safety at risk.

Moreover, Doran et al. (2017) explore different perspectives on XAI across various research fields and identify four types of XAI systems: opaque, interpretable, comprehensible, and more advanced type. It draws attention to the variations in the methods and definitions of explainability used by different AI research communities [4]. The authors argue that the current understandable and interpretable models must be revised to offer comprehensive explanations. To create AI systems that can be easily explained, Doran et al. (2017) emphasize the need for ongoing study in both interpretable and understandable systems. They also highlight the importance of integrating logic into XAI.

In addition, explainability in AI is essential for decision-makers to justify the system's outputs and processes to stakeholders, including executives, shareholders, and regulators [10]. The need for explainable AI systems grows as AI models become more complex and their decision-making capabilities more autonomous. Hoffman et al. (2019) emphasize that a good explanation should be clear, precise, and satisfying to the user. Trust in AI systems is complex, involving aspects of justified trust, unjustified trust, and mistrust. The paper explores various scales and methods to measure and maintain appropriate levels of trust, ensuring that users can confidently rely on AI systems while being aware of their limitations.

2.3 Adaptive explanations

Adaptive explanation refers to an AI system's ability to tailor its explanations based on the user's level of knowledge and information needs [23]. In the context of AI, this means the system can assess the user's familiarity with certain concepts or tools and adjust its dialogue accordingly to provide more or less detail as needed. Torrey et al. (2006) found that adaptive explanations meant that a beginner would receive detailed explanations and background information, while an expert would receive concise, technical information. This approach aimed to improve communication efficiency, user satisfaction, and task performance by ensuring that explanations were appropriately detailed for the user's expertise level.

Additionally, Han et al. (2020) recognize that preferences for robot explanations are inherently subjective and shaped by cultural and individual differences. They propose that

adaptive explanations, which are customized to an individual's preferences and cultural background, represent a critical area for future research. Although the study offers insights into general preferences for robot explanations, it emphasizes the necessity for further research to investigate how these explanations can be tailored to accommodate the diverse needs of users, thereby improving the effectiveness of human-robot interactions [9].

Similarly, Verhagen et al. (2023) explored adaptive explanations but focused on different dimensions of personalization, such as user trust, workload, and performance [26]. It demonstrates that personalized explanations can significantly enhance user satisfaction and trust in the agent. However, it also finds that explanations adapted to performance can sometimes lower task performance due to the extra time required to process detailed information. This research emphasizes the importance of developing robust user models to tailor explanations according to individual user characteristics effectively. While the studies mentioned above underline the benefits of personalization in explanations, they also highlight different aspects and challenges, underscoring the need to design and implement adaptive strategies to optimize human-agent interaction carefully. Although there are some studies on adaptive explanations, it is still unknown how adaptive explanations influence human trust and supervision during dynamic task allocation.

3 Methods

3.1 Design

We conducted an experiment to compare an agent providing progressively abstract explanations against an agent offering non-adaptive explanations (the baseline). Using a between-subjects design, participants were exposed to the adaptive or baseline/non-adaptive explanations in a simulated firefighting scenario. The study aimed to assess how adaptive explanations influenced human supervision and trust in the robotic agent. By adjusting explanations based on human cognitive sensitivity (detailed in Section 3.7), we evaluated whether adaptive explanations resulted in higher trust and explanation satisfaction compared to non-adaptive explanations.

3.2 Participants

We recruited 40 university students aged 18 to 30, comprising 19 females and 21 males. Most participants (35) were between 18 and 23 years old, while the remaining five were between 24 and 35. Regarding educational backgrounds, one participant had not completed high school, 12 had a high school diploma as their highest level of education, 16 were pursuing a Bachelor's degree but had not yet graduated, one held an associate degree, eight had completed a Bachelor's degree, and two completed their Master's degree. In terms of gaming experience, 10 participants had no gaming experience, eight had little experience, seven had a moderate amount of experience, five had considerable experience, and 10 were highly experienced gamers. All participants signed an informed consent form prior to participating in the study.

We tried to balance age, gender, education, and gaming experience equally between the baseline/non-adaptive and the adaptive agent. The baseline/non-adaptive and adaptive agent explanation conditions were homogeneous with respect to gender ($\chi^2(1) = 0$, $p = 1$). Moreover, we believe these variables might have an effect on the independent variables, so we controlled for the following variables: age, gender, education, gaming experience,

risk propensity, trust propensity, and utilitarianism. Results showed that for the control variables: age ($W = 170, p = 0.1637$), education ($W = 143, p = 0.1075$), gaming experience ($W = 231.5, p = 0.3906$), risk propensity ($W = 221.5, p = 0.5694$), trust propensity ($t(38) = -0.17524, p = 0.8618$), and utilitarianism ($W = 240.5, p = 0.2786$), there were no significant differences between the conditions. Therefore, we could exclude their influence on our measures and proceed with the analyses.

3.3 Hardware and Software

To conduct this experiment, we used a laptop and the Human-Agent Teaming Rapid Experimentation (MATRX) software, a Python package tailored for human-agent teaming research (https://matrx-software.com/). The laptop facilitated the launch and access to our two-dimensional grid world built with MATRX. Subjective measures were gathered using Qualtrics, and MATRX automatically recorded objective measures.

3.4 Environment

A dynamic task allocation system was created using MATRX to collaborate between a semi-autonomous firefighting robot and a human supervisor in a simulated firefighting environment.

- **Environment:** The 2D environment simulated various firefighting scenarios, including situations with 11 victims in need of rescue. It was designed to mirror realistic firefighting conditions, considering factors like the number of victims and how long the fire had been burning.
- **Task:** The main task for both the agent and the human supervisor was to search for and rescue victims in this simulated setting. The robot was programmed to predict how dangerous a situation was based on certain features. Depending on the level of danger it predicted, the robot would either make decisions on its own or let the human supervisor decide.
- **Agent:** The virtual firefighting robot named Brutus evaluates the danger by looking at factors such as the number of victims and the duration of the fire. Based on these evaluations, it determines whether to act on its own or to involve the human supervisor. Crucially, the human supervisor can always step in and override the robot's decisions if necessary. The robot provides explanations for its decisions, detailing how each feature influenced its assessment of danger. These explanations aim to help the human supervisor decide when to intervene or when to trust Brutus's judgment.

3.5 Task

The objective of the task was to locate injured victims in various areas and transport them to the drop zone within a 15-minute timeframe. Specifically, only firefighters could carry critically injured victims, while the agent could individually evacuate mildly injured victims. Additionally, any fallen objects could only be removed by the agent. Four decision-making situations could occur during the task: extinguish or evacuate when the robot finds mildly injured victims in burning areas; send in firefighters to rescue critically injured victims or not; send in firefighters to help locate the fire source or not; and continue or switch deployment tactics.

In the simulated firefighting scenario, six situational features are crucial for decision-making. The first feature, resistance to collapse, estimates how long the building can withstand the fire before collapsing, with values counting down in minutes until reaching zero. Secondly, the temperature indicates the internal building temperature relative to a safety threshold, with values being lower (<), close to (<=), or higher (>). Thirdly, the total number of victims in the building is known (11 victims). Fourthly, the speed of smoke spread describes how quickly smoke disseminates, with possible values being slow, normal, or fast, initially unknown, and updated upon detection. Moreover, the fire source location feature denotes whether the fire source has been located, with values as unknown (?) or found. Lastly, the distance between the victim and the fire source measures the proximity from a victim's location to the fire source, categorized as small or large, initially unknown until critically injured victims are found.

These features guide the robot and human collaborators in making informed decisions regarding safety, urgency, and resource allocation during rescue operations. Additionally, four critical decision-making situations could occur, such as the choice between offensive and defensive deployment tactics. Offensive deployment focuses on rescuing victims, while defensive deployment prioritizes extinguishing fires. The room's temperature was closely monitored to simulate a real-life scenario with fires. If the fires continued to spread and the agent and humans did not collaborate to extinguish them, the robot could not send in a firefighter to carry the critically injured victim to the safe drop zone. This temperature threshold was an essential consideration in the task. Moreover, there was a threshold of 4.1 for predicted moral sensitivity. If the predicted moral sensitivity exceeded this threshold, the agent would ask the human to step in and decide. Otherwise, the agent would make the decision independently. The seven features contributing to the predicted sensitivity were baseline moral sensitivity, the presence of fire, urgency, the number of safe victims, the speed of decision-making, the number of firefighters, and the type of tactic used. Additionally, there was an option to allocate the decision-making responsibility to the robot or the human.

3.6 Agent Types

This experiment tested two types of agents: the baseline/non-adaptive agent and the adaptive agent. Both agents performed the same tasks and were configured identically but differed in how they communicated with the human teammate. The baseline agent used a standard communication style that remained constant over time. For each of the four decision-making situations, it provided the same non-adaptive explanations to the human teammate (see Figure 1).

Unlike the baseline agent, the adaptive agent was not bound by a fixed communication style. Instead, it was designed to adjust its communication style dynamically based on the level of collaboration with the human teammate. As the agent and human teammate worked together, the agent's explanations and reasoning became increasingly abstract, reflecting a deeper level of collaboration. Both agents communicated with the human teammate using the chat box (see Figure 1).

Regarding behavior, both agents moved to the closest unexplored area and kept track of all explored areas. They were responsible for finding and rescuing victims, ensuring that any located victims were promptly assisted. Both agents also maintained detailed records of all explored areas, found victims, rescued victims, removed debris, extinguished fires, and human interventions. This comprehensive tracking was essential for maintaining an efficient search and rescue process within the experimental setup.

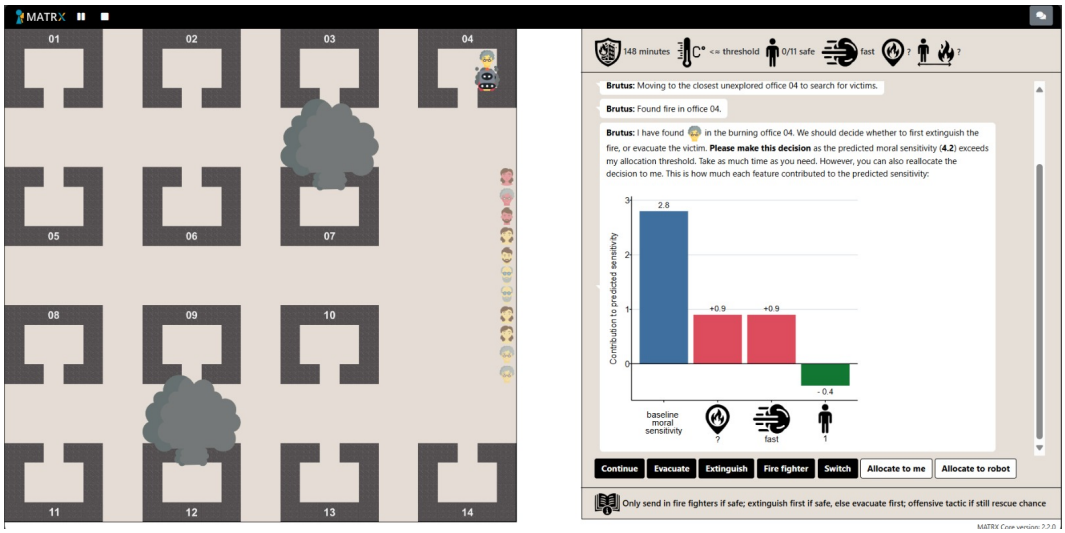

Figure 1: **Participant view of the MATRX world we used for our study**

3.7 Explanation Generation

When designing the adaptive explanations for this task, the main reason for implementing a more abstract strategy over time is the expected increase in participant experience and familiarity with the task and the explanations provided. The decision to make explanations adaptive aligns with the broader literature, which suggests that explanations should be tailored to the user's level of expertise [23]. As participants engage with the task and receive explanations multiple times, their understanding and familiarity with the task mechanics, decision-making criteria, and overall context will likely improve over time.

Consequently, as participants become more experienced, the need for detailed explanations diminishes. Detailed, granular explanations are crucial during the initial phases to ensure participants fully grasp the nuances and complexities of the task. However, as their proficiency grows, these detailed explanations can become redundant and even heavy, potentially leading to mental overload or decreased efficiency. By gradually shifting to more abstract explanations, we can maintain essential information flow while reducing cognitive load and allowing participants to focus on higher-level decision-making. This approach ensures explanations are informative and practical, capturing crucial details without overwhelming the user.

Below is an outline of a decision-making scenario during the deployment tactic (see Table 1). The textual part of the adaptive explanations was tailored for the four different decision-making situations. Table 1 illustrates what the participants read when completing the task during the deployment tactic. It is also important to note that the baseline/non-adaptive explanation only included explanations for situations with zero occurrences.

As it has been shown over time, the content of the explanation becomes more abstract and with fewer details. Each explanation includes a plot that evolves through four stages (see Figure 2, Figure 3, Figure 4, Figure 5). This approach ensures that the explanations adapt to the Brutus sensitivity level and the number of tasks completed. It is also necessary to note that the baseline/non-adaptive explanation only showed Figure 2 throughout the task.

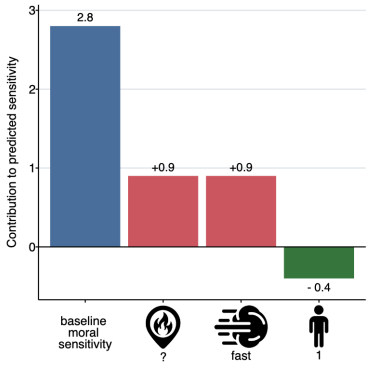

Figure 2: **Early Stage, plot with a full explanation.**

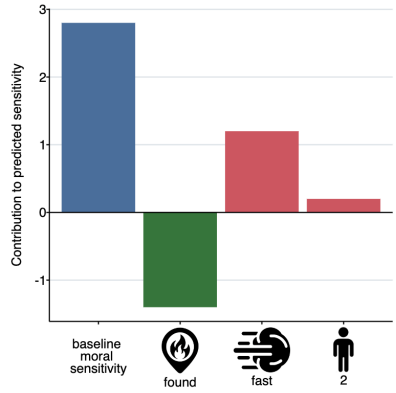

Figure 3: **Intermediate Stage, somewhat detailed.**

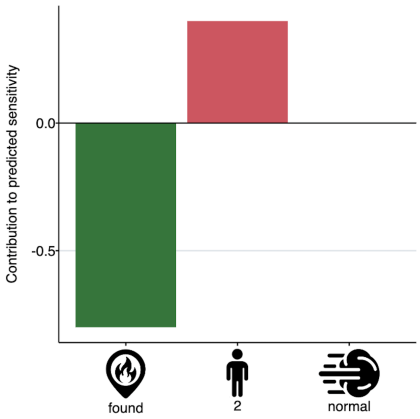

Figure 4: **Abstract Stage, less detail.**

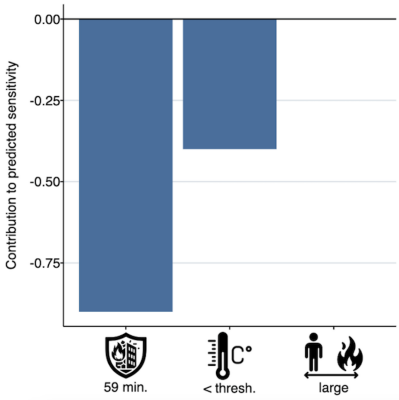

Figure 5: **Very Abstract Stage, minimal detail.**

Number of occurrences	Explanation
0	Our offensive deployment has been going on for {deployment_time} minutes now. We should decide whether to continue with this deployment or switch to a defensive deployment. **Please make this decision** as the predicted moral sensitivity (**{sensitivity}**) exceeds my allocation threshold. Take as much time as you need. However, you can also reallocate the decision to me. This is how much each feature contributed to the predicted sensitivity:
1	The current strategy has been in place for {deployment_time} minutes. A decision is needed on whether to continue or change the approach. **Your decision is required** as the predicted sensitivity (**{sensitivity}**) is beyond the threshold. You may take your time or assign this decision to me. Factors influencing the predicted sensitivity are shown below:
2	Active for {deployment_time} minutes. Continue or switch to defense? **Decision needed** due to sensitivity (**{sensitivity}**). Take your time or assign it to me. Feature contributions:
4	Decide on continuing defense or switching to offense. **Decision needed** as sensitivity (**{sensitivity}**) exceeds threshold. Take your time or assign it to me. Contributions:
6	Continue or switch to offense? **Decision needed** due to sensitivity (**{sensitivity}**). Take your time or delegate. Contributions:

Table 1: **Adaptive explanations based on the number of occurrences of the decision-making situation continue or switch deployment tactic**

3.8 Measures

We measured the dependent variables "capacity trust" and "moral trust" using a Likert scale survey [15]. Participants rated the agent Brutus on various attributes on a scale from 0 (not at all) to 7 (very). If an attribute did not fit Brutus, participants could select "does not fit." The ratings provided were used to calculate the average scores for each trust dimension. Additionally, the dependent variable "XAI satisfaction" was assessed using a Likert scale survey [10]. Participants rated their satisfaction with the explanations provided by Brutus when it allocated decision-making to either the participant or himself. The scale ranged from 1 (I disagree strongly) to 5 (I agree strongly). These responses were used to calculate the average scores for the XAI Satisfaction variable. Furthermore, the variable "disagreement rate" was measured objectively through data logged with MATRIX. This was calculated as the ratio of the number of interventions to the total number of allocation decisions.

Control variables, such as "risk propensity," were also measured using a Likert scale survey [16]. Participants rated their agreement with various statements related to their attitudes towards risk on a scale from 1 (totally disagree) to 9 (totally agree). The average of these ratings provided a mean score reflecting each participant's overall risk propensity. Similarly, "trust propensity" was measured using a Likert scale survey [17]. Participants indicated their agreement with several statements about their trust in technology, using the same scale from 1 (strongly disagree) to 5 (strongly agree). The mean score for each participant was calculated to determine their overall trust propensity, offering a subjective measure of their inclination to trust technology in general. Finally, the variable "utilitarianism" was measured using a Likert scale survey [12]. Participants indicated their agreement

with statements reflecting utilitarian ethical beliefs and values, on a scale from 1 (strongly disagree) to 5 (strongly agree). These responses were analyzed to calculate a mean score for each participant, providing a quantitative measure of their inclination towards utilitarian ethical beliefs.

3.9 Procedure

In this study, participants first opened a browser window with two tabs: one for the survey and one for the experiment. They began by reading an information sheet and providing informed consent via TU Delft Qualtrics survey. Next, participants filled in their demographic information and completed control surveys on risk propensity, the propensity to trust technology, and utilitarianism. Participants were randomly assigned to one of the two conditions: Baseline and Adaptive. For this study, some participants were assigned the baseline agent, while others were assigned the adaptive agent. They were then prompted to start a tutorial to familiarize themselves with the MATRX environment and the agent Brutus. After launching the main.py file, participants entered their ID and the explanation condition, which started the tutorial followed by the experimental task. Once the task was completed, participants returned to the survey to fill out questionnaires on trust and explanation satisfaction. All survey responses were collected using Qualtrics.

4 Results

The dependent variables, capacity trust, moral trust, XAI satisfaction, and disagreement rate, were tested to determine if there were any significant differences between the adaptive and non-adaptive explanations. The results indicated that there were no significant differences for any of these variables.

For capacity trust, the Wilcoxon rank-sum test is used because the Shapiro-Wilk normality test indicates that 'capacity trust' is not normally distributed for the non-adaptive explanation ($p = 2.716 \times 10^{-9}$). The Wilcoxon rank-sum test result ($W = 184.5$, $p = 0.6843$) suggests that there is no significant difference between the non-adaptive explanation (Mean: 5.378, SD: 0.755) and the adaptive explanation (Mean: 5.532, SD: 0.788). A p-value of 0.6843 is much higher than the common alpha level of 0.05, indicating that any observed difference is likely due to random variation rather than a systematic effect of the conditions.

Similarly, based on the Shapiro-Wilk test results, the Wilcoxon rank-sum test is used because 'moral trust' is not normally distributed for the non-adaptive explanation ($p = 0.0009988$). The Wilcoxon rank-sum test result for moral trust ($W = 178$, $p = 0.7463$) indicates no significant difference between the non-adaptive explanation (Mean: 5.315, SD: 1.699) and the adaptive explanation (Mean: 5.73, SD: 0.846). The high p-value of 0.7463 further supports this.

Furthermore, the t-test is used because the Shapiro-Wilk normality test indicates that 'xai satisfaction' is normally distributed in both conditions (baseline/non-adaptive 1: $p = 0.4794$, adaptive: $p = 0.6759$). The independent samples t-test result for XAI satisfaction ($t(38) = 0.345$, $p = 0.7319$) also shows no significant difference between the baseline/non-adaptive explanation (Mean: 3.888, SD: 0.557) and the adaptive explanation (Mean: 3.819, SD: 0.699). The p-value of 0.7319 suggests that any differences observed are likely due to chance.

Finally, the Wilcoxon rank-sum test is used because the Shapiro-Wilk normality test indicates that the 'disagreement rate' is not normally distributed in either baseline/non-

adaptive ($p = 1.973 \times 10^{-5}$) or adaptive ($p = 3.581 \times 10^{-6}$). The Wilcoxon rank-sum test result for disagreement rate ($W = 235.5$, $p = 0.2545$) indicates no significant difference between the baseline/non-adaptive explanation (Mean: 0.06, SD: 0.0899) and the adaptive explanation (Mean: 0.0285, SD: 0.0579). Although this p-value is lower than the others, it is still above 0.05, meaning that the difference is not statistically significant.

In summary, the statistical tests conducted on the four dependent variables (capacity trust, moral trust, XAI satisfaction, and disagreement rate) suggest that there are no significant differences between the baseline/non-adaptive and adaptive explanations for any of these variables. This implies that the conditions being compared do not have a notable impact on these measures.

Below are the boxplot figures (see Figure 6, Figure 7, Figure 8, and Figure 9) for all the variables mentioned above. These figures visually illustrate the distribution of the data for each condition.

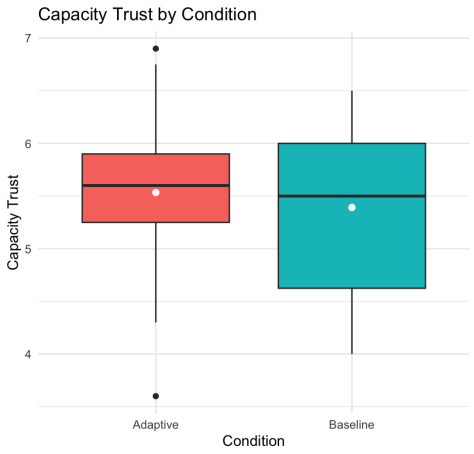

Figure 6: Capacity Trust

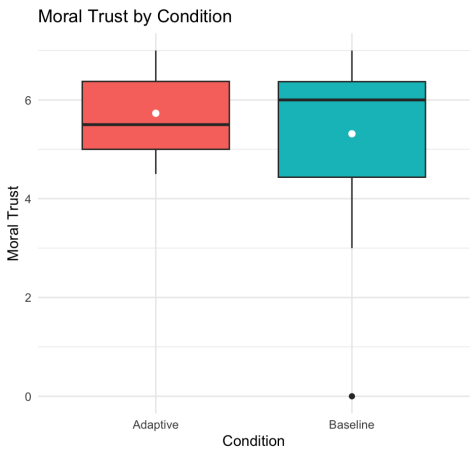

Figure 7: Moral Trust

Figure 8: XAI Satisfaction

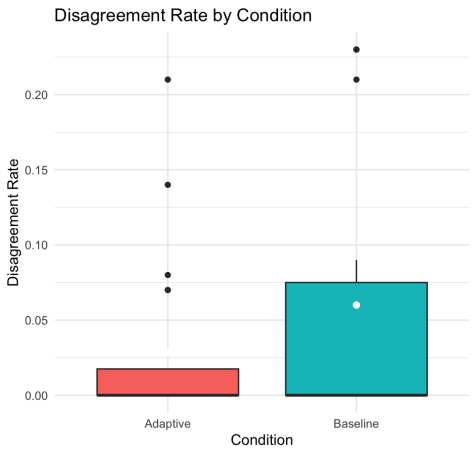

Figure 9: Disagreement Rate

5 Responsible Research

This section discusses the reproducibility of the study's methods, including the participants' selection process and any limitations related to their age. It also addresses handling participants' personal data to ensure ethical research practices. To ensure reproducibility and transparency, the study's methods are detailed thoroughly, including how participants were randomly assigned to different conditions (non-adaptive and adaptive). Participants were recruited through social media and online university channels, reaching a diverse sample. However, this method is also limited, as it may exclude individuals less familiar with or lacking access to online technologies, particularly older adults. This could introduce an age-related bias, potentially affecting the generalizability of the findings. Future studies should consider additional recruitment strategies to include a broader age range and those with limited online access. Throughout the entire study, ethical issues were not just significant, but crucial. All participants gave informed consent after being adequately informed about the study's goal and rights. In order to preserve confidentiality, participant personal data was anonymized and securely saved. In summary, this study was conducted strongly emphasizing reproducibility, ethical considerations, and the responsible handling of participant data, ensuring the research maintains its integrity and contributes valuable insights to the field. Furthermore, the code will be made public and accessible to all.[1]

6 Discussion

The results of our study indicate that there were no significant differences in the dependent variable and the control variables between the baseline/non-adaptive explanation and the adaptive explanation conditions. Specifically, the Wilcoxon rank-sum test and the t-test results showed that the differences in 'capacity trust,' 'moral trust,' 'xai satisfaction', and 'disagreement rate' were not statistically significant.

6.1 Analysis and Interpretation

The results show no significant differences in the dependent and control variables, suggesting that the type of explanation (baseline/non-adaptive) versus adaptive did not have a measurable impact on participants' responses. This outcome could be interpreted in several ways:

Firstly, both types of explanations may be equally effective in terms of influencing the measured variables. This could indicate that the choice between baseline/non-adaptive and adaptive explanations might be based on other factors, such as user preference or specific application contexts. By highlighting these factors, one can feel more engaged in the research and understand the broader implications of the findings. In addition, the research sample might have characteristics that make it less likely to show differences between the conditions. For this study, the participants were relatively homogeneous regarding their education and age range, which could reduce observed variability.

Secondly, the methods used to measure 'capacity trust,' 'moral trust,' 'XAI satisfaction,' and 'disagreement rate' may need to be more sensitive to detect subtle differences between the conditions. Despite this, it is important to note that capacity trust, moral trust, and XAI satisfaction were all quite high across both conditions. This indicates that generally,

[1]Link to the code: `https://github.com/rsverhagen94/TUD-Research-Project-2024`

participants perceived the robot as quite trustworthy and were happy with the provided explanations, which likely contributed to a low disagreement rate. The high levels of trust and satisfaction suggest that participants had a positive perception of the robot's trustworthiness and the quality of explanations, thus not often finding reasons to disagree with it. This should reassure about the effectiveness of the research and the positive impact of the explanations on the participants. Moreover, these findings might imply that the differences between the baseline/non-adaptive and adaptive explanations are insufficient to produce different outcomes in the measured variables. This could suggest that other factors, such as explanation clarity or user engagement, play a more crucial role.

Additionally, it is possible that the robot's behavior played a more significant role in influencing people's capacity and moral trust than the explanations it provided. If the robot consistently demonstrated reliable and ethical behavior, participants might have inherently trusted its actions and decisions, which would overshadow the impact of the type of explanation given. In this context, the robot's behavior could serve as a primary driver of trust, making the nature of the explanations less critical in shaping participants' perceptions. Moreover, the high levels of XAI satisfaction indicate that participants were generally quite happy with the explanations for the task allocation provided by the robot. This suggests that the baseline/non-adaptive explanations already met participants' expectations and needs. Consequently, making these explanations more adaptive over time did not add any more satisfaction. The non-adaptive explanations might have been sufficiently clear and informative, rendering additional adaptiveness unnecessary from the participants' perspective.

Futhermore, it is worth considering that the baseline explanations used in this study might not have been 'basic' enough to highlight significant differences. For example, another potential baseline explanation could have been: *I have found a victim in office {room_name}. We should decide whether to send in a firefighter to rescue the victim or if this is too dangerous.* ***Please make this decision*** *as the predicted moral sensitivity {sensitivity} exceeds my allocation threshold. I will ask for your decision after 25 seconds, but you can take as much time as you need. However, you can also reallocate the decision to me.* Completely removing the feature contributions part might have highlighted differences between the explanations more effectively. Perhaps the differences would have been more pronounced if the baseline explanations had been simplified to this extent.

Our findings differ notably from those of Torrey et al. (2006), who found that adaptive explanations significantly improved user satisfaction and task performance by tailoring the level of detail to the user's expertise [23]. In our study, no significant difference was found between adaptive and non-adaptive explanations, suggesting that the context and user population play a crucial role in determining the effectiveness of adaptive explanations. The homogeneous nature of our participant group might have minimized the perceived benefits of tailoring explanations to individual knowledge levels.

Similarly, Han et al. (2020) emphasized the subjective nature of preferences for robot explanations and the importance of cultural and individual differences [9]. While our study did not explicitly account for cultural background interpretations, the uniformity in participant demographics could explain why adaptive explanations did not show a measurable impact. This points to the necessity of considering a more diverse sample in future research to uncover the potential benefits of adaptive explanations across different cultural contexts.

Verhagen et al. (2023) highlighted that personalized explanations could enhance user satisfaction and trust but sometimes at the cost of performance due to increased cognitive load [26]. Our results, however, indicated high trust and satisfaction levels with both types of explanations, without any performance decrement. This discrepancy might be attributed to

different dimensions of personalization (e.g., workload and performance). It underscores the complexity of designing adaptive explanations and the need for comprehensive user models that account for various user characteristics and situational factors.

While our findings suggest no immediate advantage of one type of explanation over the other, understanding the nuances and contexts in which these explanations are deployed remains a critical area for future research. Emphasizing this aspect can help one sense the significance of the research's implications and the potential for further advancements in the field.

6.2 Limitations and Future Work

One limitation of the study is the recruitment method. Using social media and online university channels might have excluded older adults or those less familiar with online platforms, potentially limiting the findings' applicability to a broader population. In addition, expanding the sample size and diversity could also help detect potential differences. Including participants from various backgrounds and different levels of familiarity with the subject could provide more results. Future studies should include a more diverse demographic to understand how adaptive explanations affect different user groups.

Furthermore, it's important to note that the methods used to measure 'capacity trust,' 'moral trust,' 'XAI satisfaction,' and 'disagreement rate' may need to be more sensitive to detect subtle differences between the conditions. This underscores the need for future research to consider employing more sensitive measures or alternative methodologies, which could lead to more nuanced and accurate findings. Running experiments with a more basic/less detailed but non-adaptive baseline and experiments with higher or lower thresholds could provide additional insights into the relative effectiveness of adaptive explanations. This approach would help determine the optimal level of detail and adaptability required for different user groups and contexts. Lastly, the study focused on short-term responses to the explanations. Future work should investigate the long-term effects of adaptive explanations on user trust and satisfaction and their impact on behavior over time.

To conclude, our study offers preliminary insights. However, it's crucial to underscore the significance of addressing these limitations in future research. Doing so will deepen our understanding of the impact of different types of explanations and significantly enhance the robustness and generalizability of the findings, making the research more impactful and relevant.

7 Conclusions

This research investigated the influence of adaptive explanations that become more abstract over time on human supervision and trust in robots. Despite the theoretical benefits of adaptive explanations, our empirical results show no significant difference between baseline/non-adaptive and adaptive explanations across key dependent variables: capacity trust, moral trust, XAI satisfaction, and disagreement rate. The results also suggest that people perceived the robot doing dynamic task allocation as quite trustworthy both in terms of capacity and morality and were quite satisfied with both types of explanations supporting the allocation. The lack of significant differences suggests that making explanations more abstract over time does not necessarily enhance human supervision or trust in robots.

Future research could explore alternative forms of explanation adaptivity, different contexts, or more diverse participant groups to uncover conditions under which adaptive expla-

nations might be more beneficial. In summary, while adaptive explanations are intriguing, our study did not find evidence that they improve human supervision or trust in robots compared to baseline explanations. This suggests that more research is needed to understand how to design and implement adaptive explanations in human-robot interaction effectively.

References

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 1078 – 1088, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[2] Jose J. Canas. Ai and ethics when human beings collaborate with ai agents. *Frontiers in Psychology*, 13, 2022.

[3] Lakshita Dodeja, Pradyumna Tambwekar, Erin Hedlund-Botti, and Matthew Gombolay. Towards the design of user-centric strategy recommendation systems for collaborative human-ai tasks. *International Journal of Human-Computer Studies*, 184:103216, 04 2024.

[4] Derek Doran, Sarah Schulz, and Tarek R. Besold. What does explainable ai really mean? a new conceptualization of perspectives, 2017.

[5] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), 01 2023.

[6] Hebah Elgibreen and Kamal Youcef-Toumi. Dynamic task allocation in an uncertain environment with heterogeneous multi-agents. *Autonomous Robots*, 10 2019.

[7] Tinka Giele, Tina Mioch, Mark Neerincx, and John-jules Meyer. Dynamic task allocation for human-robot teams. volume 1, 01 2015.

[8] Megha Gupta. *Explainable Artificial Intelligence (XAI): Understanding and Future Perspectives*, pages 19–33. 11 2022.

[9] Zhao Han, Elizabeth Phillips, and Holly A. Yanco. The need for verbal robot explanations and how people would like a robot to explain itself. *J. Hum.-Robot Interact.*, 10(4), sep 2021.

[10] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2019.

[11] Matthew Johnson, Jeffrey Bradshaw, Paul J. Feltovich, Catholijn Jonker, Birna Riemsdijk, and Maarten Sierhuis. The fundamental principle of coactive design: Interdependence must shape autonomy. volume 6541, pages 172–191, 01 2010.

[12] Guy Kahane, Jim A.C. Everett, Brian D. Earp, Lucius Caviola, Nadira S. Faber, Molly J. Crockett, and Julian Savulescu. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, 125(2):131–164, March 2018.

[13] Esther Kox, Jose Kerstholt, Tom Hueting, and Peter De Vries. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, 35, 2021.

[14] Bryan Lavender, Sami Abuhaimed, and Sandip Sen. Relative effects of positive and negative explanations on satisfaction and performance in human-agent teams. *The International FLAIRS Conference Proceedings*, 36, 05 2023.

[15] Bertram F. Malle and Daniel Ullman. A multidimensional conception and measure of human-robot trust. *Trust in Human-Robot Interaction*, 2021.

[16] R.M. Meertens and R. Lion. Measuring an individual's tendency to take risks: the risk propensity scale. *Journal of Applied Social Psychology*, 38:1506–1520, January 2008.

[17] Stephanie M. Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. I trust it, but i donât know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, 55(3):520–534, 2013. PMID: 23829027.

[18] Kazuhiko Momose, Troy Weekes, Rahul Mehta, Cameron Wright, Josias Moukpe, and Thomas Eskridge. Patterns of effective human-agent teams. pages 1–13, 04 2023.

[19] R.R. Murphy. Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):138–153, 2004.

[20] M.N. Nicolescu and M.J. Mataric. Learning and interacting in human-robot domains. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 31(5):419–430, 2001.

[21] Femi Osasona, Olukunle Amoo, Akoh Atadoga, Temitayo Abrahams, Oluwatoyin Farayola, and Benjamin Ayinla. Reviewing the ethical implications of ai in decision making processes. *International Journal of Management Entrepreneurship Research*, 6:322–335, 02 2024.

[22] Fabian Ranz, Vera Hummel, and Wilfried Sihn. Capability-based task allocation in human-robot collaboration. *Procedia Manufacturing*, 9:182–189, 2017. 7th Conference on Learning Factories, CLF 2017.

[23] Cristen Torrey, Aaron Powers, Matthew Marge, Susan R. Fussell, and Sara Kiesler. Effects of adaptive robot dialogue on information exchange and social relations. pages 126 – 133, 2006.

[24] Jasper van der Waa, Jurriaan van Diggelen, Luciano Cavalcante Siebert, Mark Neerincx, and Catholijn Jonker. Allocation of moral decision-making in human-agent teams: A pattern approach. In Don Harris and Wen-Chin Li, editors, *Engineering Psychology and Cognitive Ergonomics. Cognition and Design*, pages 203–220, Cham, 2020. Springer International Publishing.

[25] Jasper van der Waa, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, and Ioana Cocu. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI*, 8, 2021.

[26] 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*, page 2316â2318, 2023. Green Open Access added to TU Delft Institutional Repository 'You share, we take care!' â Taverne project https://www.openaccess.nl/en/you-share-we-take-care Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public. ; 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS'23 ; Conference date: 29-05-2023 Through 02-06-2023.

[27] Wendell Wallach and Shannon Vallor. *Moral machines: From value alignment to embodied virtue*, pages 383–412. Oxford University Press, United Kingdom, September 2020.

[28] A. v. Wissen, Y. Gal, B. A. Kamphorst, and V. Dignum. Human - agent teamwork in dynamic environments. *Computers in Human Behavior*, 28:23–33, 2012.