Influence of Global Explanations on Human Supervision and Trust in Agent

Explainable AI for human supervision over firefighting robots

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

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

With the rise of AI presence in various contexts and spheres of life, ensuring effective human-AI collaboration, especially in critical domains, is of utmost importance. Explanations given by AI agent can be of great assistance for this purpose. This study investigates the impact of global explanations, explaining general allocation rules, on human supervision and trust in AI within the critical domain of a firefighting scenario, where human and AI agent have to collaborate to save victims. To this end, a user study involving 40 participants was performed. The user study compared the baseline and global explanation scenario and the participants' trust in the AI and explanation satisfaction were measured. The results indicated no significant differences between the two types of explanations, and in fact both achieved similar satisfactory outcomes. This suggests comparable effectiveness of global explanations in enhancing human-AI collaboration. Essentially, the insights of this study underscore the need for further exploration into contextual factors influencing the impact of global explanations and contribute to designing better human-AI teaming systems in dynamic and ethically sensitive environments.

1 Introduction

With the rapid development in the field of Artificial Intelligence (AI), the integration of artificially intelligent agents will only grow more widespread in all kinds of contexts. Even though AI agents have shown to be very capable of efficiently solving complex tasks, often times their application tends to be a "black box", i.e. mysterious or hard to understand to the user[20]. This lack of understanding of the actions of the AI agent can pose significant risks when it comes to more sensitive and critical domains, where the agent can directly impact people's well-being and thus a well-melded human-agent collaboration is crucial[12].

Existing research has explored some of the challenges of achieving effective human-agent collaboration[16]. One of the most significant ones is related to the efficient allocation of moral decision-making tasks within teams of humans and AI agents. A way to tackle this is through the design of Artificial Moral Agents (AMAs) that adhere to several Team Design Patterns (TDPs) [16]. These TDPs vary in the level of involvement and autonomy of the agent, ranging from agents assisting in decision making to fully autonomous agents making independent moral decisions [16]. Building on top of this, another challenging area of interest has to do with maintaining Meaningful Human Control (MHC), to whose end existing research has investigated how different TDPs affect human-control and decision making [17]. The findings indicate that MHC is achieved when the consequences of the human decisions can be directly observed. On the other hand, when the outcomes are delayed, perceived control was diminished.

In addition to the above-mentioned, another aspect that plays an important role in a well-integrated human-agent collaboration proves to be the usage of explanations by the agent[18]. Essentially, explanations can serve as a tool to combat the "black box" notion of AI agents and moreover the lack thereof constitutes both a practical and an ethical issue [4]. In fact, existing research has shown that the more personalized and well-suited to the situation an explanation is, the greater the benefit for the human-agent team, since higher explanation satisfaction and trust in the agent is achieved [19]. In addition, previous work also mentions that explanation must strive to be clear, timely, relevant and concise, since due to stress and big workload, humans do not have the chance to fully comprehend and interpret the results, and even seldomly read them [17]. However, there remains a knowledge gap in understanding how specific types of explanations impact human supervision and trust

in agents, especially in dynamic and uncertain environments requiring ethical judgment. Addressing this gap is essential for designing effective human-robot collaboration systems.

This research paper is concerned with a specific type of explanations, namely global explanations, in the context of a firefighting scenario where an AI agent and human collaborate to save maximum number of victims. The paper will address the question: "How do global explanations that explain general allocation rules influence human supervision over and trust in the robot?". To this end, this paper also discusses key characteristics of effective global explanations, which in essence provide a broad view of the logic behind the model, and offers comparison to the baseline scenario. Finally, this study gives insights on the circumstances under which global explanations relate to satisfactory human-robot collaboration outcomes. The motivation behind investigating specifically global explanations out of other types of explanations lies in their ability to reveal underlying principles and systematic relationships within complex models. This valuable ability can potentially lead to improved model transparency, human-AI collaboration, and robustness across diverse datasets and scenarios, which is very desirable in the field of AI.

Through performing a sufficient literature review, that led to the design and implementation of global explanations, followed by a user study and data analysis of the resulting outcomes, this paper concluded that there is no significant difference between the baseline scenario and the global explanation scenario in terms of trust and satisfaction. However, both types of explanations achieved similar high results, which implies comparable effectiveness of global explanations in human-AI collaboration. The following sections explain in detail the process.

2 Background

2.1 Key Characteristics of Global Explanations

Global explanations identify and describe general patterns or trends in the model[2]. Their aim is to explain the behavior of the model in its entirety, unlike local explanations that are aimed at explaining individual decision and instances[3]. In essence, global explanations provide a broad view of the logic behind the model and can even be seen as summaries of more complex explanations[13]. As such, they offer much more scalability compared to other types of explanations. In contrast, the information that local explanations convey regarding the agent's decision-making is state-specific [6], as is the case with the baseline explanations in this study, made clear in section 3.6.

Existing work on generation of global explanations involves various machine learning methods, of which most commonly used are decision trees, rule-based classifiers and neural networks [13]. Another popular way of constructing global explanations is through combining local explanations based on SHapley Additive exPlanation (SHAP) values [8]. The SHAP values provide a unified measure of feature importance by computing the average contribution of each feature across all possible combinations of features [9]. In this way, the most influential features can be identified and used in the making of global explanations.

These approaches among others are mainly applied in the context of actual black-box models. However, in this research paper, the focus is not so much on the generating process, rather than the actual effect of global explanations on the user. Thus, the model and its intricacies are known for the design of this study and this knowledge is taken advantage of when creating the global explanations, as described in section 3. Tools like the abovementioned decision trees and SHAP values, despite their usefulness, prove to be unnecessary,

since the features that influence the model globally are known in advance.

2.2 Research Question Description

The research question this work aims to answer is: "How do global explanations that explain general allocation rules influence human supervision over and trust in the robot?". The answer to this question will be reached through comparison to the baseline explanation scenario following the methodology described below.

3 Method

3.1 Experiment Design

To investigate the influence of global explanations on human supervision and trust in robot a user study was performed. The study was structured to compare two scenarios - baseline explanations scenario and global explanations scenario. Thus, it had a 2x1 between-subjects design with the scenario type as the independent variable. As dependent variables, we measured perceived capacity trust and moral trust in the AI agent, as well as satisfaction with the explanations and disagreement rate, i.e. the amount of times the participants disagreed with the robot's decision and allocated it to themselves.

3.2 Participants

A total of 40 participants were recruited for this study, with 20 participants allocated to the baseline scenario group and another 20 participants assigned to the global explanations scenario group. Out of the total sample size, 37.5% were female, and 62.5% were male. More specifically, of the participants appointed to the baseline scenario, 45% were female and 55% were male, while of the participants selected for the global scenario, 30% were female and 70% were male. A Chi-square test for homogeneity was performed for gender to determine if there are statistically significant differences in the distribution of gender across the two different scenarios or the distribution is similar. The results ($\chi^2 = 0.42667$, $df = 1$, $p = 0.5136$) showed that the proportion of males and females is consistent across the two scenarios, indicating that gender is homogeneously distributed and any observed differences in outcomes are not due to gender distribution discrepancies. The vast majority of the participants fell within the age range of 19 to 24 years old, 95% to be exact, while the remaining 5% were withing the age range of 24 to 34. Only one participant was above 24 in the samples of both the baseline and the global scenario. In terms of education, all participants possessed at least a high school diploma, 52.5% had only high school diploma, 27.5% had some college credit, 12.5% had Bachelor's degree and the rest 7.5% had Master's degree. For the baseline scenario, 45% fell into the first category (high school diploma), 40% were of the second category (some college credit), 5% were of the third (Bachelor's degree) and 10% had Master's. For the global scenario, the respective percentages in the same order were 60%, 15%, 20%, and 5%. Concerning gaming experience, there was a lot of variety among the participants, but most had at least a little experience. The exact percentage values for the total sample size were: 15% of all participants had no gaming experience at all, 15% had a little experience, 22.5% had moderate, 17.5% had a considerate experience and finally 30% had a lot of experience. For the baseline scenario, the percentage values in the same order were: 25% with no experience, 10% with a little, 20% with moderate,

15% with a considerate experience and finally 30% had a lot of experience. For the global scenario the percentage values in the same order were: 5%, 20%, 25%, 20%, 30%. For the above-mentioned variables of age, education and gaming experience a two-samples Wilcoxon test was performed in order to assess whether there is a statistically significant difference between the distributions of these variables across the two scenarios. The results indicate no significant difference for all three - age ($W = 200$, $p = 1$), education ($W = 218$, $p= 0.6039$) and gaming experience ($W = 176.5$, $p = 0.5233$). Additionally, data on risk propensity, trust propensity, and utilitarianism of each participant was collected. Section 3.8 provides a detailed explanation of how these variables were measured. For risk propensity, trust propensity, and utilitarianism an independent samples t-test was attempted (if the data assumptions were met) since it has more statistical power and it is more sensitive to detecting differences between groups. However, if the assumptions were not met, a Wilcoxon test was performed. The results indicated no significant difference for all three variables - risk propensity ($W = 200$, $p = 1$), trust propensity ($t = 1.6153$, $df = 36.996$, $p = 0.1147$), and utilitarianism ($W = 252$, $p = 0.1621$). Informed consent was obtained from all participants before their participation in the study. Ultimately, none of the previously described variables (age, gender, education, gaming experience, risk propensity, trust propensity, utilitarianism) required further control during the data analysis stage, as all of them were similar and did not have significant differences across the two scenarios.

3.3 Hardware and Software

The required hardware for the experiment described in section 3.1 is a laptop (or desktop computer) capable of running Python and the MATRX software, a Python package designed to facilitate human-agent teaming research (https://matrx-software.com/). The laptop was used to launch a two dimensional grid world created using MATRX. Additionally, a stable internet connection is needed for conducting the survey through Qualtrics, which is a web-based platform used for creating and distributing surveys, as well as collecting and analyzing survey data.

3.4 Environment

The environment used in this research is a MATRX world, depicted in Figure 1. The world features 14 designated areas that represent offices in a building, some of which on fire, with smoke spreading out and victims trapped inside. To ensure diverse victim representation, four victim types (woman, man, older woman, older man) with two injury categories (critical and mild) were included, denoted by victim color (red for critical, yellow for mild). The AI agent is depicted in the upper right corner in Figure 1, and is essentially a firefighting robot. Additionally, there is a human agent (controlled by the participants) that communicates with the robot agent through a chat window, illustrated in Figure 2. On the right side of Figure 1, there is a designated safe zone where the victim drop-off happens. It is also worth mentioning that the world is made to resemble an actual realistic firefighting situation with the help of 6 situational features (observed at the top of the chat window in Figure 2) that change as the time progresses, namely temperature level, smoke spread, resistance to collapse, number of victims, fire location source and distance between victim and fire source. In this way, the model strives to come close to reality.

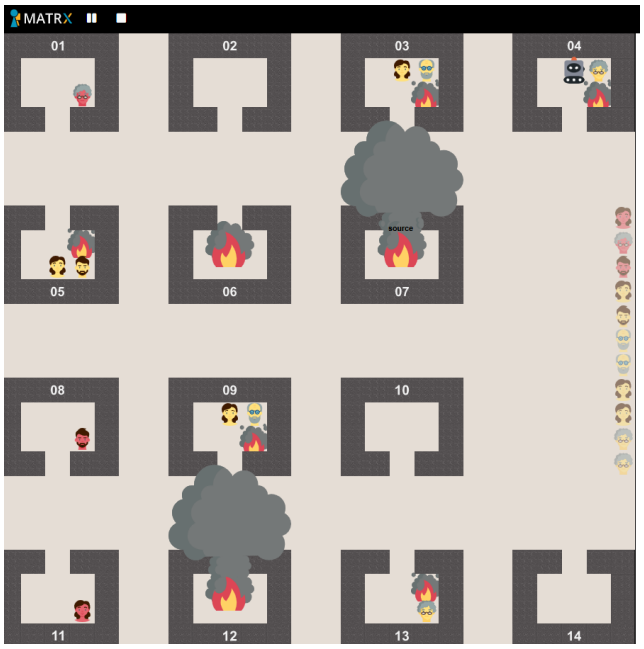

Figure 1: Screenshot of the model environment and scenario used in this study.

3.5 Task and Agents

As already mentioned, the given scenario is a firefighting situation in a burning building where some of the offices are on fire and there are people trapped inside. The objective is to achieve the best possible outcome for both the firefighters and the people trapped inside, some of which more injured than others, with the help of an AI agent that is capable of making decisions based on its surrounding circumstances.

The robot (AI agent) is tasked with the job of performing search and rescue operations in collaboration with a human agent(in this case the user of the system). The robot sustains an ongoing communication with the human through a chat window (depicted in Figure 2), explaining its actions. Whenever the robot encounters situations that are deemed morally sensitive above a certain threshold (for example whether to evacuate the people or extinguish the fire first), it asks the human to make the decision. The human agent then can make a decision by clicking one of the buttons at the bottom of the chat window in Figure 2. Other than that, the robot works independently, informing the human of its actions during the whole time. The human is capable of intervening at any moment and command the course of action of the robot, if he/she chooses to.

3.6 Global Explanation Generation

There are 4 decision-making situations that can happen and that require additional justification and possibly intervention from the human - choosing deployment tactic, choosing whether to evacuate or extinguish first, choosing whether to locate the fire source, and choosing whether to rescue critically injured victims (is the environment safe enough for the firefighters). All of these decisions and their allocation depend on the predicted moral

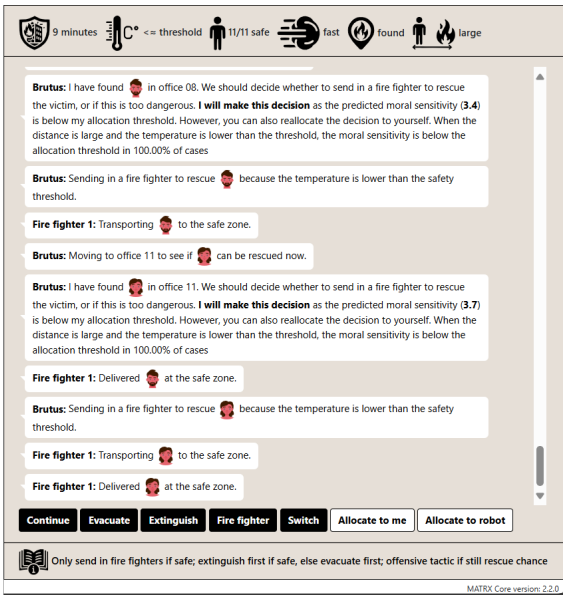

Figure 2: Screenshot of the chat window used for communication between the agent and the robot.

sensitivity of the robot. If that sensitivity is above the threshold, the decision is allocated to the human. The sensitivity can be calculated in 4 different ways depending on the circumstances, and thus there are 4 different sensitivity functions. Each depends on different set of the 6 situational features mentioned above in section 3.5. Table 1 gives more details on these dependencies. The exact way that the functions calculate the moral sensitivity can be found in the publicly accessible code repository used for this study[1].

Sensitivity Function	**Temperature**	**Smoke spread**	**Resistance**	**Victims**	**Fire Location**	**Distance**
Function 1	yes	-	yes	-	-	yes
Function 2	yes	-	yes	yes	-	-
Function 3	-	yes	-	yes	yes	-
Function 4	-	-	yes	yes	yes	-

Table 1: Dependency of Sensitivity Functions on Situational Features

Before delving into the implementation details of global explanations, it is important to first discuss the baseline explanations, as they form the foundation against which all comparisons are made. The baseline explanations are also presented in the 4 decision-making situations, introduced earlier. They are quite rich in detail and frequently display visually how each of the features contributes to the final result of moral sensitivity in the specific situation. With this in mind, they can be considered local explanations, since they are tailored for a specific instance and do not capture the general logic or trends of the model. Figure 3 illustrates what a typical baseline explanation looks like.

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

Situational Feature	Range
Temperature	[Lower, Close, Higher]
Smoke spread	[Slow, Normal, Fast]
Resistance	[0 to 150]
Victims	[None, Unclear, One, Multiple]
Fire Location	[Known, Unknown]
Distance	[Small, Large]

Table 2: Situational Features and Their Ranges

Back to the general explanations, by examining the already-mentioned 4 moral sensitivity functions and trying out a variety of values within the ranges of the features used as parameters, one can make observations about possible dependencies between the parameters and the value of the outcome of the moral sensitivity. Table 2 depicts the possible values of each of the 6 features used as input for the moral sensitivity functions. All possible values within each parameter's range were tried out and plugged in their corresponding sensitivity function. Observing the resulting value of moral sensitivity and calculating the percentages of the cases when the result is below or above the threshold enabled obtaining general statements depicting the overall logic of the model. These statements were essentially used as the building blocks of the global explanations. Below as an example are presented 2 global explanations for each sensitivity function, the last sentence in bold is the general statement explained above.

- Function 1
  - I have found (`victim-type`) in office (`office-number`). We should decide whether to send in a firefighter to rescue the victim, or if this is too dangerous. I will make this decisionas the predicted moral sensitivity (`sensitivity-value`) is below my allocation threshold. However, you can also reallocate the decision to yourself. **When the distance is small and the temperature is higher than the threshold, the moral sensitivity is below the allocation threshold in 76.92% of cases.**
  - I have found (`victim-type`) in office (`office-number`). 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-value`) exceeds my allocation threshold. Take as much time as you need. However, you can also reallocate the decision to me. **When the distance is large and the temperature is higher than the threshold, the moral sensitivity is above the allocation threshold in 30.77% of cases.**
- Function 2
  - The fire source still has not been located. We should decide whether to send in fire fighters to locate the fire source, or if this is too dangerous.I will make this decision as the predicted moral sensitivity (`sensitivity-value`) is below my allocation threshold. However, you can also reallocate the decision to yourself. **When there are no people involved and the temperature is close to the threshold, the moral sensitivity is below the allocation threshold in 100.00% of cases.**

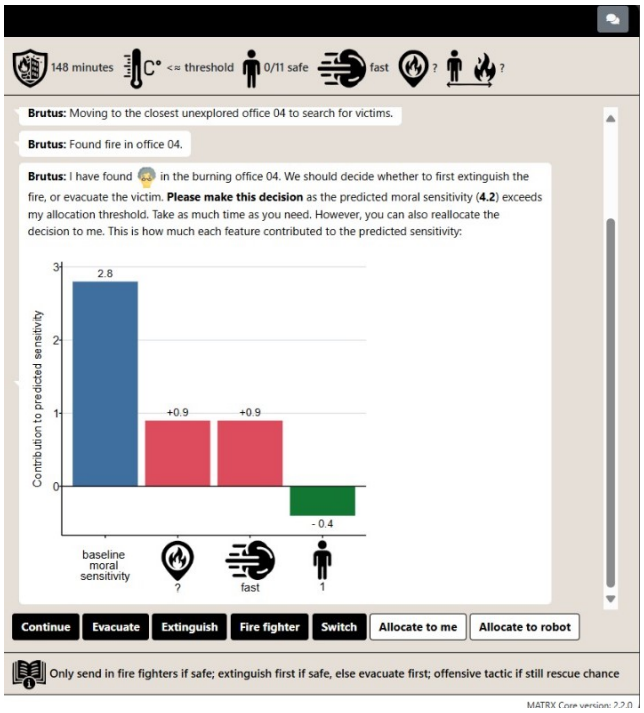

Figure 3: An example of a baseline explanation.

  - The fire source still has not been located. We should decide whether to send in fire fighters to locate the fire source, or if this is too dangerous.Please make this decision as the predicted moral sensitivity (`sensitivity-value`) exceeds my allocation threshold. Take as much time as you need. However, you can also reallocate the decision to me. **When multiple people are involved and the temperature is close to the threshold, the moral sensitivity is above the allocation threshold in 100.00% of cases.**

- Function 3
  - I have found (`victim-type`) in the burning office (`office-number`). We should decide whether to first extinguish the fire, or evacuate the (`victim-type`). I will make this decision as the predicted moral sensitivity (`sensitivity-value`) is below my allocation threshold. However, you can also reallocate the decision to yourself. **When the number of victims is 1 and the smoke spreads fast and the location is known, the moral sensitivity is below the allocation threshold in 100.00% of cases.**
  - I have found (`victim-type`) in the burning office (`office-number`). We should decide whether to first extinguish the fire, or evacuate the (`victim-type`). Please make this decision as the predicted moral sensitivity (`sensitivity-value`) exceeds my allocation threshold. Take as much time as you need. However, you can also reallocate the decision to me. **When the number of victims is 1 and the smoke spreads fast and the location is unknown, the moral**

sensitivity is above the allocation threshold in 100.00% of cases.

- Function 4
  - Our offensive deployment has been going on for (`offensive-deployment-time`) minutes now. We should decide whether to continue with this deployment, or switch to a defensive deployment (`additional-explanation`). I will make this decision as the predicted moral sensitivity (`sensitivity-value`) is below my allocation threshold. However, you can also reallocate the decision to yourself. **When there are no people involved and the location is unknown, the sensitivity is below the allocation threshold in 100.00% of cases.**
  - Our offensive deployment has been going on for (`offensive-deployment-time`) minutes now. We should decide whether to continue with this deployment, or switch to a defensive deployment (`additional-explanation`). Please make this decision as the predicted moral sensitivity (`sensitivity-value`) exceeds my allocation threshold. Take as much time as you need. However, you can also reallocate the decision to me. **When there are multiple people involved and the location is known, the sensitivity is above the allocation threshold in 100.00% of cases.**

The examples given above serve as illustration of what the global explanations that were implemented in the system look like. All global explanations can be found in the public code repository[2], mentioned earlier.

3.7 User Study Procedure

As already mentioned, 20 participants were recruited for the baseline scenario, and 20 participants were recruited for the global explanations scenario. The procedure, that each one of them followed, is straightforward and is described below.

1. **Consent Form**: Each participant is asked to read information about the study and asked to read and sign a consent form detailing the study's purpose and procedures.
2. **Demographic Survey**: Each participant is asked to answer a number of demographic questions to collect information such as age, gender, and background.
3. **Risk Propensity Evaluation**: Each participant is asked to answer questions related to risk-taking behavior, in order to assess the participant's tendency to take risks.
4. **Trust Propensity Evaluation**: Each participant is asked to answer questions related to trust in technology in various contexts, in order to assess the participant's inherent tendency to trust technology.
5. **Utilitarianism Scale**: Each participant is asked to answer questions related to utilitarian ethical beliefs and values, in order to assess the participant's inclination towards utilitarian ethical reasoning.
6. **Introduction**: Each participant is introduced to the model described in section 3.4 and given time to get familiar with it by playing a tutorial version of the experiment.

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

7. **Global Explanation Version**: Each participant plays out the game in the global explanation version. After the time is up or after all victims are saved, the participant continues with the survey.

8. **MDMT**: The Multi-Dimensional Measure of Trust - Each participant is asked to fill out questions about the AI agent according to the standard MDMT consisting of 16 items that assess four differentiable dimensions of trust. One can trust an agent because the agent is Reliable, Capable, Ethical, and/or Sincere.

9. **Explanation Satisfaction Evaluation**: Each participant is asked to indicate his/her degree of satisfaction with the presented explanations in the scenario.

3.8 Measures

As mentioned in the previous section 3.7, firstly the control variables were measured. Demographic data, which includes age, gender, education level and gaming experience, was collected by having participants select the options that they most identified with. Risk propensity was assessed with the Risk Propensity Scale developed by Meertens and Lion [10], while trust propensity was evaluated using a Propensity to Trust scale proposed by [11]. Finally, utilitarianism, was measured with the Oxford Utilitarianism Scale (OUS), which consists of the Impartial Beneficence subscale and the Instrumental Harm subscale, by [7]. Following, after each participant finished playing out the experiment, multiple logs summarizing their activity were generated. These logs give insights about the completeness of the game, the allocations of the robot and human agents, as well as total interventions by the human and the disagreement rate. Afterwards, the participants filled out the MDMT form [15], from which capacity trust and moral trust are measured. If an attribute from the form did not fit the robot, participants could choose the "does not fit" option. These ratings were then utilized to obtain the mean score for each of the two trust dimensions. Finally, XAI (explainable artificial intelligence) satisfaction was measured using [5]. Participants evaluated their satisfaction with the explanations given by the robot using a scale ranging from 1 (strongly disagree) to 5 (strongly agree). These ratings were then used to compute the mean scores for the XAI Satisfaction variable.

4 Responsible Research

Reflecting on the ethical aspects of this research, several measures were taken to ensure the study was conducted responsibly and ethically. Firstly, all participants signed an informed consent and were provided with sufficient information about the study's purpose, procedure, any potential risks and the analysis of personal data. This ensured that all participants were fully aware of their involvement and made an informed decision about their participation and the collection of their data. Secondly, regarding sensitive information, the privacy and confidentiality of the participants was also ensured by anonymizing and secure storage of their identification data. Only aggregate data were reported in the study, ensuring that individual responses could not be traced back to any specific participant. Considering the study design itself, efforts were made to minimize any potential harm or discomfort to participants. The questionnaires and tasks were designed to be non-invasive and respectful of participants' time and well-being. Each participant was free to withdraw from the study at any point in time.

Regarding the reproducibility of the methods, detailed description of the study procedures, questionnaires, and analysis techniques has been maintained. In this way, other researchers can replicate the study under similar conditions to verify the findings or explore related questions. The use of standardized measures such as the Multi-Dimensional Measure of Trust (MDMT) and established scales for risk propensity,trust propensity and utilitarianism further facilitate the reproducibility and reliability of the study, as all of these tools have been validated in prior research. In addition, to enhance reproducibility and transparency even more, the code repository used for both scenarios was made publicly accessible[3].

5 Results

As previously mentioned in section 3, in order to investigate the effects of global explanations on human supervision and trust in robot, several variables were measured during and after the experiment. This includes variables such as capacity trust, moral trust, XAI satisfaction, disagreement rate, completeness, total allocations, human allocations, total interventions. The variables of particular interest are namely capacity trust, moral trust, XAI satisfaction, disagreement rate, which are the dependent variables of this study. For each of these variables an independent samples t-test was attempted, however none of them met the data assumptions. Therefore, a two-samples Wilcoxon test was performed. Interestingly, none of the four variables showed any significant differences between the two scenarios: capacity trust ($W = 169.5$, $p = 0.4166$), moral trust ($W = 163$, $p = 0.3231$), XAI satisfaction ($W = 174.5$, $p = 0.4975$), disagreement rate($W = 230.5$, $p = 0.329$). Figure 4 shows the box plots illustrating the capacity trust and moral trust of the baseline scenario (condition 1) compared to the global explanation scenario (condition 4). Figure 5 shows the box plots illustrating the XAI satisfaction and disagreement rate of the baseline scenario (condition 1) compared to the global explanation scenario (condition 4).

Figure 4: Box plot of the capacity trust and moral trust of the baseline scenario (condition 1) vs. the global explanation scenario (condition 4).

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

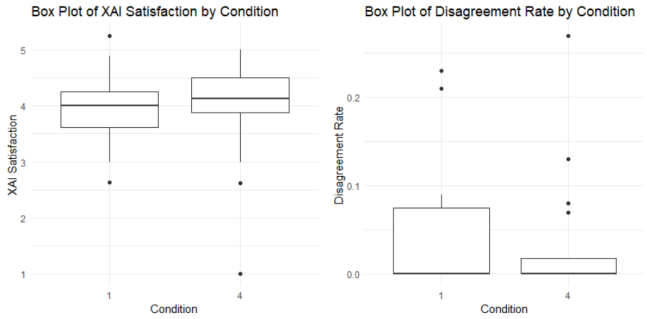

Figure 5: Box plot of the XAI satisfaction and disagreement rate of the baseline scenario (condition 1) vs. the global explanation scenario (condition 4).

6 Discussion

6.1 Result Interpretation and Analysis

The present study explores the effects of global explanations on human supervision and trust in robots by examining variables such as capacity trust, moral trust, XAI (explainable artificial intelligence) satisfaction, and disagreement rate. As mentioned in section 5, no significant differences were found between the global explanations scenario and the baseline explanation.

Looking at the box plot for capacity trust (Figure 4 on the left), the median values for the baseline condition (Condition 1) and the global explanation condition (Condition 4) are very similar, suggesting that global explanations did not have a big and noticeable impact on participants' trust in the robot's capacity. The interquartile ranges (IQR) for both conditions are also quite comparable, which implies that the spread of trust levels is similar across the different scenarios. The presence of outliers in the global scenario indicates some variability of the collected data, however, as mentioned in section 5, the Wilcoxon test results ($W = 169.5$, $p = 0.4166$) confirm that there is no statistically significant difference in capacity trust between the two conditions.

In similar fashion, the box plot for moral trust (Figure 4 on the right) shows that the median values are close for both conditions. An evident difference is the size of the IQR for the baseline condition, which is noticeably larger than the one of the global explanations, suggesting greater variability in moral trust ratings for the baseline condition. Additionally, outliers in both conditions indicate extreme values in some participants' responses. However, as already described in section 5, any differences between the baseline and the global explanations scenarios were not significant as demonstrated by the results of the Wilcoxon test ($W = 163$, $p = 0.3231$).

Regarding XAI satisfaction, Figure 5 shows very similar median values and IQRs for both condition. Both conditions have outliers, indicating variability in satisfaction levels. Despite this, the Wilcoxon test results ($W = 174.5$, $p = 0.4975$) show no significant difference in XAI satisfaction between the baseline and global explanation conditions.

Figure 5 also depicts the disagreement rate measurements for the two conditions, that visually differ only in the size of the IQRs. For the baseline scenario, the IQR is wider, which indicates larger variability in disagreement rates among participants compared to the

global scenario. Nevertheless, the Wilcoxon test ($W = 230.5$, $p = 0.329$) indicated a lack of significant differences between the two conditions.

All of the above leads to the conclusion that the participants' capacity trust, moral trust, XAI satisfaction and disagreement rate were not significantly affected by the global explanations. Global explanations had no effect at large, since they did not substantially change the median levels of trust in the robot's capacity or morality, nor did they meaningfully impact XAI satisfaction or disagreement rate. These results can be interpreted in multiple ways.

One way of interpretation is that the regular baseline explanation are equally as expressive and powerful as the global ones. The effects of the baseline and the global explanations on the dependent variables is highly comparable. With this in mind, perhaps making the decision to use one over the other lies in other aspects of the situation, such as the goal within the context, the task complexity, system scalability, personal preference or specific application in a certain domain. In fact, global explanations are considered better suited for the purpose of system scalability since they can handle larger complexity and scale of data as they facilitate generalization and provide a complete mental model of the systemâs workings [14].

The fact that both (local) baseline explanations and global explanations were perceived relatively equally well in our scenario is in line with the findings of Aechtner et al. [1]. This research has explored whether so-called AI novices (people with little to no AI knowledge) prefer local over global explanations generated by local and global XAI methods respectively[1]. The research reveals that there is no indication of a preference for local over global explanations and in fact the mean score for explanations generated by global methods is slightly greater than for local methods in terms of usefulness, informativeness and overall satisfaction which comprise the evaluation criteria [1].

However, in another research by Sivaprasad et al. [14], it is claimed that global model explanations perform better at providing understandability and building trust compared to local model explanations. Our findings differ since the results cannot show a clear dominance of global explanations over the baseline local explanations when it comes to trust. A possible reason might be found in the composition of the experimental sample. The participants involved in the user study exhibited a lot of similarities when it comes to age and education, which could have possibly decreased the variability in the outcomes. Another possible reason behind the lack of clear differences in the dependent variables lies in the measurements themselves. The way, that the variables of capacity trust, moral trust, XAI satisfaction and disagreement rate were measured, might not have been sufficiently sensitive to detect more nuanced differences.

Nevertheless, the dependent variable for both conditions all have very satisfactory values. It is important to highlight that particularly capacity trust, moral trust and XAI satisfaction were consistently high, which suggests that the robot was viewed as capable, moral, trustworthy and able to elaborate its actions well. This is the case for both conditions, which might pose the questions of how basic the baseline explanations actually were and how much more satisfaction making these explanations more global could have actually added. Indeed, the baseline explanations were quite informative and rich initially and perhaps making them simpler (but still local/non-global) could have resulted in more distinct differences in terms of XAI satisfaction. Moreover, it is plausible that the high scores for the dependent variables of capacity and moral trust can be attributed to the actions of the robot during the scenario even more so than the explanations that the robot gave. During the gameplay, the robot acted independently rather frequently and was capable and consistent in its tasks with no

attempts to do something malicious or unethical. As a result, the robot would naturally win over the trust of the participants, disregarding the effect of the type of explanations, which would lead to a very satisfactory evaluation of the robot in terms of trust afterwards no matter baseline or global.

To sum up, the results of this study indicate that there is no significant differences in capacity trust, moral trust, XAI satisfaction and disagreement rate. Despite this, interpreting and analyzing this outcome and the possible factors leading up to it is fundamental for future research and advancement in the field of Explainable AI.

6.2 Limitations and Future Work

Regarding the limitations of this study, the sample size of this user study and its composition were constrained to a certain extent. This constraint might have increased the likelihood of Type II errors, where actual differences between baseline and adaptive explanations could go undetected. Future research should strive for a larger number of participants and a greater diversity among them, as this would possibly better facilitate the generalization of the findings and make the detection of differences easier. Another limitation to consider would be the methodology of measuring the variables. As already mentioned in the previous section 6.1, the way that the dependent variables are measured could be modified such that it captures more subtle differences. Additionally, the value of the moral sensitivity threshold mentioned in section 3.5, that causes the robot to either allocate the decision to himself or to the human agent, could be changed to potentially lead to more insights about the effect of different types of explanations. Similarly, as briefly stated in section 6.1, a revision of the baseline explanations might be also helpful for discovering more clear differences and pinpoint the effect of global explanations more precisely. Furthermore, a different approach to generating the global explanations could be considered. Future research should consider these factors to better examine the effect of global explanations and make further advancements in the field.

7 Conclusions

This research studied the effect of global explanations on human supervision and trust in AI agent within a firefighting scenario. Key characteristics of effective global explanations were identified, emphasizing the importance of clarity, relevance, and conciseness. Those characteristics include the depiction of global trends and the general logic behind the model. A structured user study involving 40 participants was performed to compare the baseline explanations to the global explanations scenario. The results did not indicate any significant differences in the measured dependent variables between the two scenarios, which include capacity trust, moral trust, XAI (explainable artificial intelligence) satisfaction and disagreement rate. In fact, the measurements were very similar and quite high for both conditions. This suggests that both types of explanations were perceived equally well by participants and both provided a comprehensive overview of the AI's decision-making process, leading to satisfactory human-robot collaboration outcomes. Even though the findings of this research cannot show a decisive advantage of one type of explanation over the other, understanding the factors and contexts, in which these global explanations were deployed, highlights the importance and relevance of this study and paves the way for further inquiry on this matter. Future research could explore the applicability of global explanations across more diverse

user populations and different scenario to examine its effectiveness to a greater extent. Additionally, further studies could investigate the integration of global explanations with other types of explanations or how global explanations compare to other kinds of explanations such as contrastive, on-demand and adaptive, which can contribute to advancements in the field of AI.

References

[1] Jonathan Aechtner, Lena Cabrera, Dennis Katwal, Pierre Onghena, Diego Penroz Valenzuela, and Anna Wilbik. Comparing user perception of explanations developed with xai methods. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, 2022.

[2] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020.

[3] Guillermo Fernandez, Riccardo Guidotti, Fosca Giannotti, Mattia Setzu, Juan A. Aledo, Jose A. Gámez, and Jose M. Puerta. Flocalx - local to global fuzzy explanations for black box classifiers. In I. Miliou, N. Piatkowski, and P. Papapetrou, editors, *Advances in Intelligent Data Analysis XXII. IDA 2024. Lecture Notes in Computer Science, vol 14642*, pages 197–209. Springer, Cham, 2024.

[4] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. 51(5):93, 2018.

[5] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023.

[6] Tobias Huber, Katharina Weitz, Elisabeth AndrÃ©, and Ofra Amir. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, 301:103571, 2021.

[7] 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, 2018.

[8] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. Explainable ai for trees: From local explanations to global understanding, 2019.

[9] Wilson E. MarcÃlio and Danilo M. Eler. From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347, 2020.

[10] Ree M. Meertens and René Lion. Measuring an individual's tendency to take risks: the risk propensity scale 1. *Journal of Applied Social Psychology*, 38(6):1506–1520, 2008.

[11] 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.

[12] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.

[13] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Glocalx - from local to global explanations of black box ai models. *Artificial Intelligence*, 294:103457, 2021.

[14] Adarsa Sivaprasad, Ehud Reiter, Nava Tintarev, and Nir Oren. *Evaluation of Human-Understandability of Global Model Explanations Using Decision Tree*, page 43â65. Springer Nature Switzerland, 2024.

[15] Daniel Ullman and Bertram F. Malle. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, pages 618–619, 2019.

[16] Jeroen van der Waa, Jurriaan van Diggelen, Lukas Cavalcante Siebert, Mark Neerincx, and Catholijn 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 19â24, 2020, Proceedings, Part II*, pages 203–220. Springer International Publishing, 2020.

[17] Jeroen van der Waa, Sanne Verdult, Kyra van den Bosch, Jurriaan van Diggelen, Tamara Haije, Stefan van der Stigchel, and Ingrid Cocu. Moral decision making in human-agent teams: Human control and the role of explanations. 8:640647, 2021.

[18] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Ranjay Krishna, Tobias Gerstenberg, and Michael S. Bernstein. When do xai methods work? a cost-benefit approach to human-ai collaboration. In *CHI '22: Conference on Human Factors in Computing Systems*, pages 1–15, New York, NY, USA, 2022. ACM.

[19] Ruben S. Verhagen, Mark A. Neerincx, Can Parlar, Marin Vogel, and Myrthe 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*, pages 2316–2318, 2023.

[20] Carl Zednik. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34:265–288, 2021. Received 11 March 2019; Accepted 14 October 2019; Published 20 December 2019; Issue Date June 2021.