Explainable AI for human supervision over firefighting robots

The influence of on-demand explanations on human trust

Elena Negrila
Supervisors: Myrthe Tielman, Ruben Verhagen
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 Negrila
Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Ruben Verhagen, David Tax

Abstract

In human-AI agent interactions, providing clear visual or textual explanations for the agent's actions and decisions is crucial for ensuring successful collaboration. This research investigates whether having the visual explanations displayed only on-demand, instead of having them consistently shown as the baseline, has an impact on the human supervisor's level of confidence and satisfaction with the AI agent. Therefore, a case study of 40 participants was conducted to explore this hypothesis and the participants were divided into 2 groups, one interacting with the on-demand condition, and the other with the baseline one. Through questionnaires, the participants' capacity and moral trust in the robot, the explainable artificial intelligence satisfaction, and the disagreement rate with the robot's decisions have been collected. Demographic data was gathered from the participants to explore whether their background could impact the collaboration. This data included the participants' gender, age, education, gaming experience, risk propensity, trust propensity, and utilitarianism. The resulting statistical analyses indicated no significant differences between the baseline and the on-demand conditions concerning trust and explanation satisfaction. This suggests that the overall collaboration was not primarily impacted by the frequency of visual explanations requested on demand. Although the results implied a high satisfaction with the interaction, further studies with more diverse user groups are recommended. Overall, this research reinforces the importance of transparency in decision-making processes during the collaboration between an AI agent and a human supervisor.

1 Introduction

Collaborations between humans and autonomous agents span various domains [1]. In this research, their collaboration was studied concerning the firefighting domain. Even though robots are becoming progressively self-reliant, the tasks implying moral decisions are considered extremely sensitive [2]. Therefore, it is still advised to have human supervisors actively involved in the process of collaboration, as they perform better in unexpected situations [3].

The firefighting domain presents some dangerous aspects, requiring quick decision-making and the ability to adapt to hazardous environments [4]. Robots, with their ability to perform repetitive tasks with high speed and having a lower safety risk than humans, present an invaluable asset for search and rescue situations [5]. They can perform a variety of tasks, such as navigating through smoke-filled buildings, evacuating victims, and extinguishing fires [6, 7]. However, the involvement of human collaborators remains crucial, especially for making decisions in situations with a high moral sensitivity, which the robots are not yet equipped to properly handle [2].

In this collaborative process, the optimal form of interaction remains uncertain. To better understand and meet human needs, various types of robot explanations should be explored. As semi-autonomous robots begin to make more intricate and critical decisions, it becomes imperative to ensure that their choices are reliable and moral. Ethical decision-making requires actions to be justifiable with clear explanations. As such, for humans to trust the agents, it is essential to understand the ethical frameworks they employ and to ensure that these are applied consistently and predictably [8].

Therefore, one thing to take into consideration is whether humans respond better to visual explanations compared to textual ones, and if visual aids should be presented only at specific key moments during the collaboration. This research aims to explore how the possibility of requesting additional on-demand explanations influences explanation satisfaction and the level of trust of the human supervisor in the robot. In the baseline version of the collaboration, the visual statistics describing the rationale behind the decision are displayed at each decision point of the robot. The on-demand explanations build on top of the baseline, and the visuals are only displayed upon the human supervisor's request. This feature is particularly relevant in high-stakes environments like firefighting, where having the right balance between information overload and understanding the rationale behind a robot's actions can be crucial for making appropriate informed decisions.

Exploring this research direction works towards better insights into the collaboration between robots and human supervisors. More specifically, this research aims to identify strategies that can enhance the collaborative efficacy between robots and human supervisors. It represents a step further in the development of a trustworthy AI-human communication system [9].

The main research question can be divided into subquestions to achieve structured reasoning and a better understanding of the specific goals of this research, while also facilitating data collection. The subquestions derived from this topic and further explored are:

1. How do the on-demand explanations differ from the baseline explanations?
2. How often do users require on-demand explanations from the robot?
3. Does the background of the users impact how frequently they require explanations?
4. Is the frequency of the request for additional information correlated to the level of trust in the robot?
5. Do the users still need more insight, even after receiving this extra information?

2 Background of the research

2.1 Human-Agent Teamwork

Human-agent collaboration combines the precision and speed of autonomous agents with the capacity of humans to make morally sensitive decisions. For instance, in the firefighting environment, robots can navigate through smoke-filled buildings, identify fire sources, and locate victims to gather critical data. The human supervisor contributes to strategic decision-making and moral judgments. This type of teamwork improves safety by allowing robots to perform dangerous tasks and reducing the exposure of human firefighters.

Dynamic task allocation systems enable robots to adjust their tasks based on continuously changing environmental conditions. However, formal analysis tools play an important role in those systems [10]. Robots can provide real-time data and situational assessments, to ensure that human firefighters have the necessary information to prioritize tasks effectively. For example, deciding whether to first extinguish fires or evacuate victims calls for dynamic task allocation. Combined with a time-pressure situation, this enhances the level of implication of the human supervisor in the task, as they can make decisions themselves and bear the responsibility for them [11].

In the context of firefighting, meaningful human control ensures that semi-autonomous robots can effectively contribute without compromising safety or ethical standards. However, human supervisors maintain the ability to supervise and intervene at any point in the decision process. Some studies on the traceability of meaningful human control have observed a positive impact on the transparency of the semi-autonomous system [12]. Making the decision process of the agents accessible to human supervisors, through visual or textual explanations has been shown to enhance trust and accountability [13].

2.2 Explainable Artificial Intelligence

Explainable AI (XAI) is essential in helping human collaborators understand AI decision-making models [14]. In situations where life-or-death decisions need to be made quickly, transparency and accountability provided by XAI are crucial for effective collaboration. Understanding the precise reasoning behind the decision of an AI agent can enhance the effectiveness of firefighting operations. When human supervisors can see that the robot's decisions are ethically justified, they are more likely to trust and approve these actions [8].

Previous research indicated that the level of trust in the automation process is closely linked to its performance [15]. This trust can be significantly enhanced through XAI, which provides clear and understandable insights into the AI agent's decision-making process. In morally sensitive situations, trust in the AI's decisions is crucial for gaining human collaborators' approval. Trust has a significant impact on the approval of the agent's decisions, especially in morally sensitive situations. Therefore, through its transparency, XAI can provide an effective collaboration in human-AI agent teams.

Additionally, XAI enhances the effectiveness in a firefighting environment by executing and explaining complex data analyses and predictions. Ensuring transparency within AI systems is crucial for maintaining accountability, especially in the event of errors. By following the AI's decision-making processes, human supervisors can identify and rectify mistakes, ultimately improving overall safety. Robot agents can help locate and extinguish fires considering factors such as temperature, fire spread speed, and fire intensity [7]. XAI provides clear and understandable data, helping firefighters prioritize rescue operations and extinguishing efforts based on solid reasoning.

However, providing explanations that are both accurate and easily understandable to human supervisors, who may not have technical backgrounds, remains a challenge. Previous research indicates a gap between delivering explanations to users without domain experience and personalizing these explanations to the user's context [16].

In this research, baseline explanations in human-robot interactions include visual graphs at each decision point of the robot. These graphs explain the factors considered, such as smoke spread, available time before building collapse, distance to victims, and temperature. Building on these baseline explanations, the focus of this thesis is on-demand explanations, which are tested to observe if they enhance trust and explanation satisfaction.

2.3 On-demand explanations

Human-agent collaboration has been approached in other works, where it has been observed that personalized explanations from a robot agent improve the associated level of trust [13]. These personalized explanations are tailored to focus on agents who use a user model of the involved human to personalize their explanations. However, a question to be answered is whether on-demand explanations, where humans can explicitly request additional information, which is not shown otherwise, can improve the trust of the human collaborator. Compared to the personalized explanations mentioned above, the on-demand explanations focus not on the type of information, since no user model is employed here, but on the availability of information, as they allow the human supervisor to decide when they want to receive such additional information.

Exploring this approach is particularly important in dynamic and high-stakes environments such as firefighting. Research indicates that visual representations significantly enhance situational analysis and improve the interpretability of the decision-making process [17]. This work explores whether the supervisor's ability to choose when to utilize visualizations allows for more determined and reliable decision-making processes. This capability potentially equips human supervisors to better monitor, evaluate, and trust the robot's actions.

In the context of human-robot interaction, previous studies have shown that the high quality of the interaction increases the perceived intelligence of the robot [18]. This paper introduces an additional interaction mechanism, allowing human supervisors to request visualization of specific statistics when needed. This flexibility ensures that visual aids are employed only when deemed necessary by the supervisor, enhancing the relevance and effectiveness of the information provided, while also reducing potential information overload on the supervisor.

This section has discussed various types of explanations and methods to enhance human-agent collaboration, supported by prior studies focusing on personalized agent explanations. The following sections analyze whether providing additional information only upon human request improves the human supervisor's trust in the robot. This investigation is performed by implementing on-demand explanations and comparing them to the existing baseline through a user study where participants act as human collaborators.

3 Method

3.1 Design

The design included a user study with 40 participants who were divided into two groups: those interacting with the baseline condition of the explanations and those using the version with the on-demand condition. The purpose of the user study was to investigate the best means of interaction between the human supervisor and the robot agent. The responses of the participants who interacted with the on-demand condition were compared against the ones of those interacting with the baseline version.

3.2 Participants

To evaluate the efficacy of this approach, a user study was performed, where participants recruited from the personal network undertook the role of the human agent collaborating with the robot in the simulated environment. The sample size of participants based on which the feedback is created is 40, where half of the individuals interacted with the baseline condition and the other half with the on-demand condition. Before they participated in the study, all members gave their informed consent, approved by the ethics committee of our institution (ID 4002). For the considered control variables, the participant's distribution is as follows: for the identified gender (19 females, and 21 males), for the age (32 between 18-24 years old, 6 between 25-34 years old, 2 between 44-54 years old), for the previous education (13 with high-school degree, 9 with some college credit, 11 with Bachelor's degree, 6 with Masters's degree and 1 with PhD's degree) and lastly for the gaming experience (10 with no experience, 4 with little experience, 7 with moderate experience, 9 with considerable experience, and 10 with a lot of experience).

To evaluate the effectiveness and influence of the explanations, several tests were conducted on the collected data. For this purpose, several control variables have been used, namely the demographic and background factors that might influence the dependent variables and the participants' responses when interacting with the robot agent. The control variables that have been tested are gender, age, education, gaming experience, risk propensity, trust propensity, and utilitarianism. The risk propensity has been measured according to previous research [19], using a scale from 1 to 9, for the 7 questions of the questionnaire, where 1 means total disagreement and 9 total agreement, then by calculating their mean. The trust propensity [20] and the utilitarianism [21] have also used measuring techniques according to previous research. A scale from 1 to 5 was used to measure the participants' response, where 1 means strong disagreement and 5 means strong agreement, then the mean of the results was calculated. The questionnaire details can be found in the Appendix A.

In order to examine if there was a significant difference between the participants' gender and the condition of the user study they took part in, a chi-square test was conducted. No statistically significant difference between the gender distributions across the baseline and on-demand conditions has been observed (Chi-Square Statistic (χ^2): 0, p-value: 1.0, df(degrees of freedom): 1). Therefore, the randomization process concerning gender has been effective.

The two-sample Wilcoxon test was used for the age, education, and gaming experience to assess whether there was a significant difference between the distributions of the two independent samples. The results for the age (p-value = 0.01875) and for the education (p-value = 0.004427) variables both have a p-value less than the significance level of 0.05. This suggests a statistically significant difference in the age and education distributions between participants in baseline and on-demand conditions, the data being not homogeneous concerning demographic variables. For further examination of this imbalance of age and education for the baseline and the on-demand conditions, linear regression tests were run. The age and education were tested against each dependent variable, capacity trust (p-value = 0.07205 for age and p-value = 0.6397 for education), moral trust (p-value = 0.4383 and p-value = 0.6494), xai satisfaction (p-value = 0.2082 and p-value = 0.5854), and disagreement rate (p-value = 0.571 and p-value = 0.1201). Therefore, age and education are not significant predictors for the dependent variables. For the gaming experience, the results using the two-sample Wilcoxon test (p-value = 0.8897) show that there is no statistically significant difference between the conditions.

For the remaining control variables, the independent samples t-test is used if the data assumptions are met and if the data is normally distributed. Otherwise, the two-sample Wilcoxon test is used. For the risk propensity variable, the Wilcoxon test has been used and resulted in p-value = 0.001461. For a deeper investigation of this imbalance, linear regression tests were run. This resulted in p-value = 0.9204 for capacity trust, p-value = 0.1877 for moral trust, p-value = 0.9565 for xai satisfaction, and p-value = 0.03554 for disagreement rate. Therefore, risk propensity is not a significant predictor for the capacity and moral trust dependent variables, but it is for disagreement rate. For the trust propensity, no significant differences between distributions were found when running the independent samples t-test (p-value = 0.2855). In the case of utilitarianism, no significant differences have been found either when running the Wilcoxon test (p-value = 0.2335).

3.3 Hardware and Software

The experiment was conducted on a commercial local machine to launch the two-dimensional simulation of the firefighting environment. The MATRX tool (Human-Agent Teaming Rapid Experimentation Software) was employed to create and manage this environment. Qualtrics [1] was used to obtain participation consent from the participants and to collect pre- and post-interaction data through surveys. This platform facilitated efficient and standardized data collection and ensured that participants' consent was properly recorded. Excel files were used to log the data obtained from the participants and to keep track of the values for each variable. For data analysis, RStudio was utilized to perform various statistical tests for the data analysis.

3.4 Environment

Within the simulated 2D firefighting environment (see Figure 1), the collaboration between a robot and a human agent has already been established using a semi-autonomous robot interacting with a human supervisor. The two agents collaborate during critical decision-making points within the environment. This environment consists of 14 distinct areas, each potentially containing victims needing evacuation, fires to be extinguished, substantial amounts of spreading smoke, and fallen debris during the extinguishing process. There are 11 victims in total who need to be transported to the designated drop zone.

To help prioritization during the decision-making, victims are colored based on their condition: critically injured victims are marked in red, while mildly injured victims are marked in yellow. During the rescue phase, the robot is capable of transporting mildly injured victims to the drop zone, as depicted by the list of victims on the right side of Figure 1. Firefighters can also be summoned to the environment to locate fire sources or rescue critically injured victims.

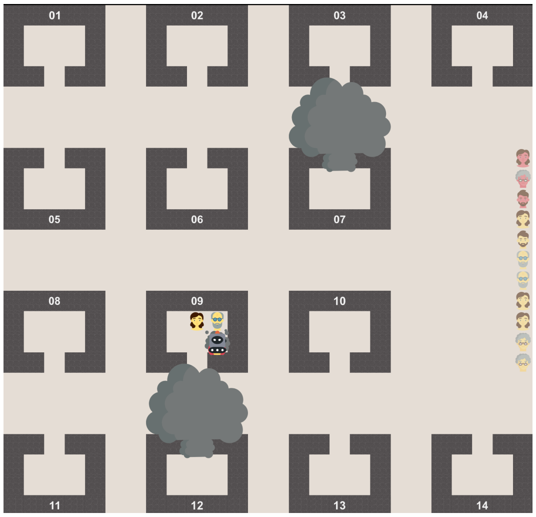


Figure 1: Firefighting virtual environment

[1] https://tudelft.eu.qualtrics.com/

3.5 Task

The objective of the task in which the human supervisor and the robot collaborate is to evacuate the victims and deliver them to the designated drop zone within the time frame. There are six situational features to take into consideration during the task: the resistance to collapse, the temperature, the number of victims, the spread of the smoke, the fire source location, and the distance between the victim and the fire source. They can be observed in Figure 4, at the top of the screen. The first one estimates the time limit on this task before the building collapses. The second represents the temperature in each area and during the decision-making process, and it is compared to a pre-established safety threshold. The next one represents the number of victims that have been rescued out of the total number of victims. This is followed by a representation of the spreading of the smoke, which can either be at a slow, normal, or fast pace. This value is updated according to each area where fire and smoke are found. The next one represents the fire source location, which can be labeled either as unknown or as found. At the beginning of the task, it is considered to be unknown. Next, there is the distance between the victim and the fire source, and it can be labeled either as small or large. This, too, is considered unknown at the beginning of the task and it is further refined when the robot finds victims that are critically injured.

There are four types of decision-making situations that the robot might face. Firstly, the choice of either employing the offensive or the defensive deployment tactic. The first tactic focuses on searching for victims and evacuating them, while the second one focuses on finding fires and extinguishing them. The decision consists of either continuing with the current tactic or switching to the alternative one. Secondly, the robot can be faced with the decision of either evacuating a mildly injured victim or extinguishing the fire in the area first. The guidelines in the robot's implementation indicate to first extinguish unless the smoke is spreading at a fast pace and the source of the fire has not yet been found. If the human supervisor makes the decision, the choice of first evacuating the victim can be expressed by pressing the 'Evacuate' button at the bottom of the screen represented in Figure 4. Otherwise, the 'Extinguish' button can be pressed. Thirdly, a decision can consist of either sending in firefighters to locate the fire source if the situation is considered safe enough or choosing not to send them. In the case where the human makes the decision, the firefighter can be sent by pressing the 'Fire fighter' button at the bottom of the screen represented in Figure 4. Otherwise, the 'Continue' button can be pressed. Finally, the robot can not evacuate the critically injured victim on its own, so a decision needs to be made on whether to send firefighters to rescue the victim, or to wait until the conditions for the firefighters to intervene no longer pose a high risk. The rescue task is chosen by pressing the 'Fire fighter' button, otherwise, the 'Continue' button can be pressed.

When interacting with the on-demand condition, the human supervisors also have the option to press the 'Extra info' button, which can be observed at the bottom of Figure 4. This displays the additional information, represented by a graph visualization of the situational features.

On both the baseline and the on-demand conditions, the robot automatically decides in situations that fall below a certain threshold of predicted moral sensitivity. However, human collaborators can intervene and allocate the decision-making to themselves, by pressing the 'Allocate to me' button presented at the bottom of Figure 4. When the moral sensitivity exceeds the threshold, the human supervisor is asked to take over, making the final decisions based on the robot's assessments. They can also re-allocate the decision-making to the robot if they choose to, by pressing the 'Allocate to robot' button.

3.6 Agent Types

The collaboration included two types of agents, the semi-autonomous robot, called Brutus, and the human supervisor. Each participant in the user study undertook the role of the human agent.

The robot was able to perform tasks such as evacuating mildly injured victims and delivering them to the drop zone, extinguishing fires, removing possibly fallen debris, searching the rooms for victims or fires, keeping track of the rescued victims and of the areas that have not yet been searched and report the tasks and the decision-making outlines to the human supervisor. However, Brutus can not deliver the critically injured victims on its own. Those victims can only be delivered by a firefighter AI agent. The robot and the human agent can communicate through the chat box presented in Figure 4, and the human agents can respond during decision-making stages by using the designed buttons.

The robot can perform dynamic task allocation by adjusting the tasks and decisions based on the interventions of the human, it can offer data analysis to the supervisor, and it can handle the execution of moral decisions.

3.7 Explanation generation

During the interaction between the robot agent and the human, the explanation generation reveals the underlying reasons behind the informed decisions of the robot agent.

At each action point during the firefighting simulation, the robot agent generates reports to be sent to the human supervisor. These reports can either detail the robot's current action or highlight morally sensitive situations requiring immediate decisions.

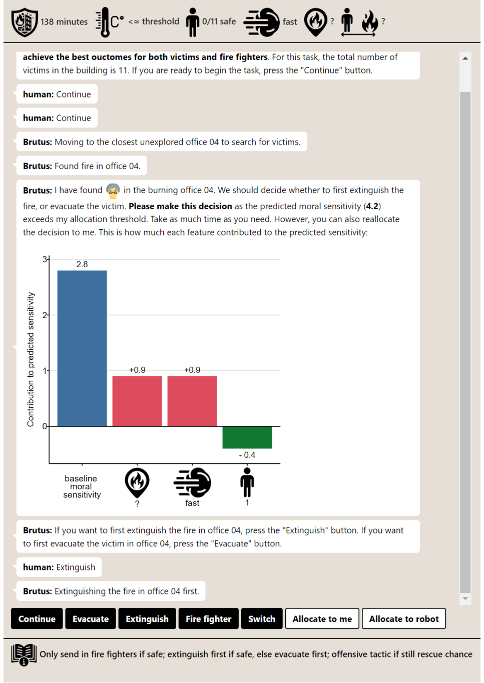
This research examines both the baseline and on-demand conditions. In the baseline condition, visual explanations are automatically presented to the human supervisor at every decision point, as shown in Figure 2. These visual graphs include baseline moral sensitivity and situational features such as the fire source location, smoke spread speed, and the number of people needing evacuation from the robot's current location.

The on-demand condition, in contrast, focuses on flexibility and responsiveness rather than constant transparency. In this condition, graph visualizations are displayed only when explicitly requested by the human supervisor, as illustrated in Figure 3. The rationale behind removing the constant visual explanations and offering them on demand is to reduce information overload and allow humans to focus on the current task and request more detailed information themselves. This way, we can investigate how the level of trust and satisfaction of the user are influenced by the on-demand explanations. This approach aims to determine if the graphs enhance the understanding of the robot's decision-making rationale and whether they should be displayed continuously or only when needed. This ensures that detailed and possibly overwhelming visual information is only shown on demand, thereby offering comprehensive data while avoiding the overburden of information.

3.8 Measures

This research measures the following dependent variables: capacity trust, moral trust, explanation satisfaction, and disagreement rate. They were measured in the questionnaire the participants were presented with after interacting with the platform.

The capacity and the moral trust were measured using the multi-dimensional measure of the trust survey [22]. The participants' trust scores were measured using a scale from 0 to 7, where 0 indicated complete disagreement with the given statement about trust, and 7



achieve the best outcomes for both victims and fire fighters. For this task, the total number of victims in the building is 11. If you are ready to begin the task, press the "Continue" button.

human: Continue

human: Continue

Brutus: Moving to the closest unexplored office 04 to search for victims.

Brutus: Found fire in office 04.

Brutus: I have found 🧓 in the burning office 04. We should decide whether to first extinguish the fire, or evacuate the victim. **Please make this decision** as the predicted moral sensitivity (**4.2**) 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:

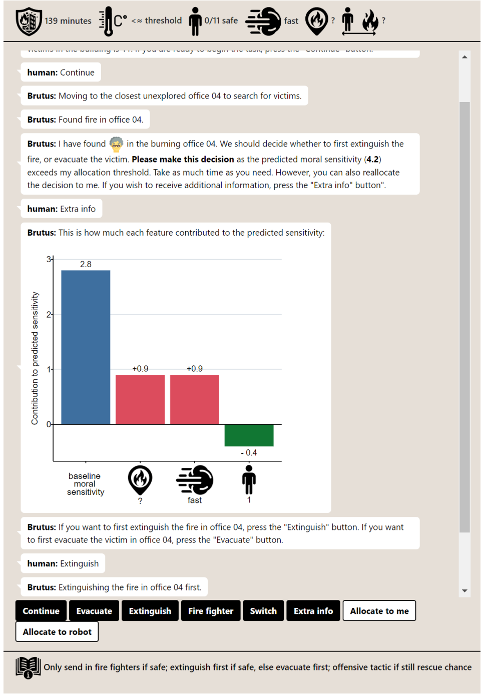
Brutus: If you want to first extinguish the fire in office 04, press the "Extinguish" button. If you want to first evacuate the victim in office 04, press the "Evacuate" button.

human: Extinguish

Brutus: Extinguishing the fire in office 04 first.

Only send in fire fighters if safe; extinguish first if safe, else evacuate first; offensive tactic if still rescue chance

Figure 2: Baseline Condition



human: Continue

Brutus: Moving to the closest unexplored office 04 to search for victims.

Brutus: Found fire in office 04.

Brutus: I have found 🧓 in the burning office 04. We should decide whether to first extinguish the fire, or evacuate the victim. **Please make this decision** as the predicted moral sensitivity (**4.2**) exceeds my allocation threshold. Take as much time as you need. However, you can also reallocate the decision to me. If you wish to receive additional information, press the "Extra info" button".

human: Extra info

Brutus: This is how much each feature contributed to the predicted sensitivity:

Brutus: If you want to first extinguish the fire in office 04, press the "Extinguish" button. If you want to first evacuate the victim in office 04, press the "Evacuate" button.

human: Extinguish

Brutus: Extinguishing the fire in office 04 first.

Only send in fire fighters if safe; extinguish first if safe, else evacuate first; offensive tactic if still rescue chance

Figure 3: On-demand condition

Figure 4: Human-Agent collaboration interface

indicated complete agreement. Additionally, there was an option labeled 'does not fit,' which participants could select if they felt a particular attribute was not applicable for describing the robot. Each type of trust was assessed with 8 questions, and the final trust score for each type was calculated by calculating the mean of the responses.

The explainable satisfaction was measured through similar approaches as those used in previous research on user satisfaction in human-AI agent interactions [5]. Participants' scores were calculated using a scale from 1 to 5, where 1 indicated strong disagreement with the presented statement and 5 indicated strong agreement. The participants responded to 8 statements and then their final satisfaction was calculated by taking the mean of their responses.

The disagreement rate was automatically logged at the end of the collaboration, by taking into consideration the number of reallocations to the human supervisor or robot.

3.9 Procedure

Before being introduced to the interface, participants completed a questionnaire designed to assess their risk propensity, trust propensity, and utilitarianism opinions. They first interacted with a trial version of the application to familiarize themselves with the platform. Following this, they proceeded with the experimental version and then, lastly, they completed an additional predefined questionnaire based on their experience as human supervisors.

Their feedback, which is integrated into the results section, focuses on their interactions with the robot. Responses to the questionnaire were logged using Qualtrics. Objective data regarding participants' interactions with the interface were automatically recorded upon task completion.

The task was deemed successful after the 40 participants tested the experimental version of the application in the user study. This feature served as a measure for analyzing the trust of the human supervisor in the robot. Each user study session lasted approximately 40 minutes and was conducted either in person or online.

4 Results

Various tests were performed on the gathered data to assess the effectiveness and impact of the explanations. This section presents the tests performed on the dependent variables: moral trust, capacity trust, XAI satisfaction, and disagreement rate, to investigate potential differences based on the on-demand explanations. The collected data is divided into two groups: one from participants who interacted with the baseline condition and the other from those who experienced the on-demand condition.

The independent samples t-test was used if the data assumptions were met. Otherwise, the two-sample Wilcoxon test was used. The significance level of the p-value against which the tests were run is 0.05. The Shapiro-Wilk test was performed on each dependent variable. If at least for one of the conditions, baseline or on-demand, the p-value was lower than 0.05, a Wilcoxon test was run. Conversely, if for both of the conditions, the p-value was equal to or greater than 0.05, an F-test to compare variances for the two conditions, followed by a t-test was performed.

For the capacity trust the mean was higher for the on-demand condition (mean = 5.661) than for the baseline (mean = 5.378), indicating a slightly higher score for the on-demand explanations. The standard deviation (sd) was similar (sd = 0.7553725) for on-demand and (sd = 0.8090469) for the baseline. The F-test to compare variances yielded a p-value of 0.7678, following the assumption of equal variances for the subsequent t-test. The two-sample t-test resulted in a p-value of 0.2604 and 38 degrees of freedom (df). The 38 degrees of freedom suggest a large sample size for analyzing differences between the conditions. This is calculated by subtracting 2 from the total number of data collections across the two groups. The sample size was 40, therefore the degrees of freedom were 38. The degrees of freedom impact the t-distribution and therefore the value for determining the significance.

For moral trust, the Wilcoxon test was performed and resulted in W = 171.5 and p-value = 0.8036. The mean moral trust score in the on-demand condition (mean = 5.782632) is higher compared to the baseline condition (mean = 5.314737). The standard deviation in the baseline condition is larger (sd = 1.698893) than in the on-demand condition (sd = 0.9135209), so there is a larger variability in the scores gathered from the baseline condition.

For XAI satisfaction, the Wilcoxon test resulted in W = 174.5 and p-value = 0.4978. The mean XAI satisfaction score in the on-demand condition (mean = 3.9975) is higher compared to the baseline condition (mean = 3.888). The standard deviation in the baseline condition is smaller (sd = 0.5570184) than in the on-demand condition (sd = 0.6894458), so scores obtained from the on-demand condition vary more.

For the disagreement rate, the Wilcoxon test resulted in W = 153.5 and p-value = 0.183. The mean disagreement rate score in the on-demand condition (mean = 0.1055) is higher compared to the baseline condition (mean = 0.06). The standard deviation in the baseline condition is smaller (sd = 0.08991224) than in the on-demand condition (sd = 0.1263027), but they can be considered fairly similar.

For all of the dependent variables, since the p-value obtained exceeds the significance threshold of 0.05, it indicates that there are no statistically significant differences between the baseline and on-demand conditions.

To compare the data distributions for the baseline and the on-demand condition, the box plot from Figure 5 is used to visualize it. The line inside the box represents the median value of the data when sorted in ascending order. The box represents the interquartile range and the whiskers extending from the boxes suggest how much the data varies in the upper and lower quartile. The individual points represented above or below the whiskers are the data outliers.

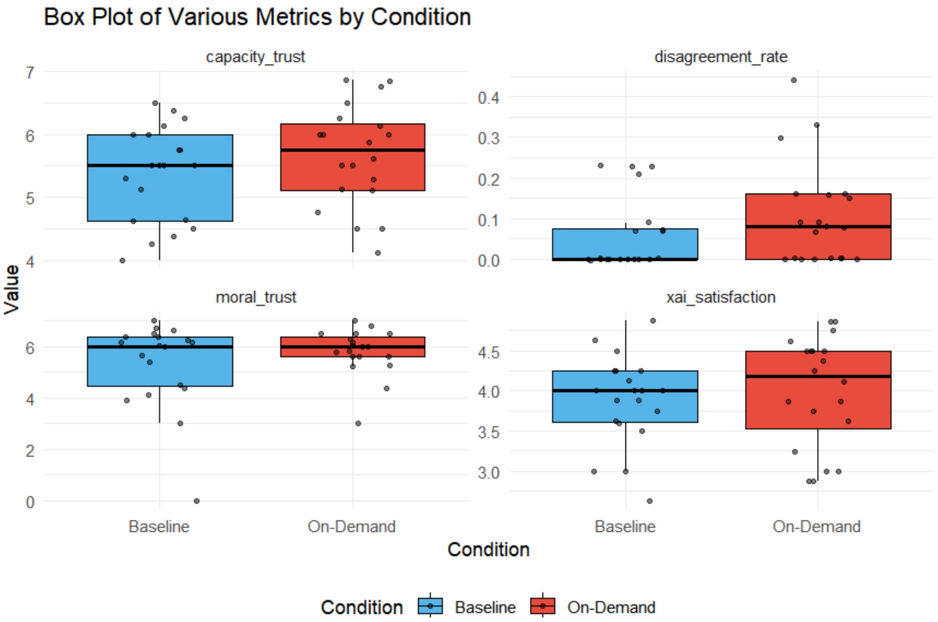


Figure 5: Data distribution in baseline VS on-demand.

For the on-demand condition, additional tests have been performed to assess the overall satisfaction of the participants. This aimed to investigate the correlation between the number of times the participants asked to be provided with additional visual graphs during decision-making and their overall satisfaction. The most noteworthy results have been obtained for capacity trust (r = 0.27 and p = 0.2542) and for moral trust (r = 0.27 and p = 0.2571), where there is a weak positive correlation, but not a significant one. Those results can be observed in the plot from Figure 8. The r value suggests a slight inclination toward higher trust when on-demand explanations are required more often. For the XAI satisfaction (r = 0.06 and p = 0.8074) and the disagreement rate (r = -0.08 and p = p = 0.7518), the results show that the dependent variables do not correlate with the number of requests for additional information. The p-values have been compared to the 0.05 significance level, where a value larger than 0.05 indicates that the results are not statistically significant. The correlation coefficient, r, measures the strength of the correlation between the measured variables.

To observe if there is any correlation between the background of the participant and the frequency of the request for an on-demand explanation, additional correlation tests were run for the control variables risk propensity, trust propensity, and utilitarianism. For the risk propensity, the results (r = -0.13 and p = 0.5800) indicate that the frequency of additional information requests slightly decreases as risk propensity increases. For the trust propensity (r = 0.07 and p = p = 0.7670) and the utilitarianism (r = 0.25 and p = 0.2972), the results indicate that the frequency of additional information requests slightly increases when these control variables increase. However, for all of these control variables, the relationship is not statistically significant, because of the p-value being higher than 0.05, and could be due to random chance. All of the correlation analysis plots for both the dependent and the control variables can be observed in Appendix A. In general, there was a low request for additional information, with the mean value of this request being 3.5.

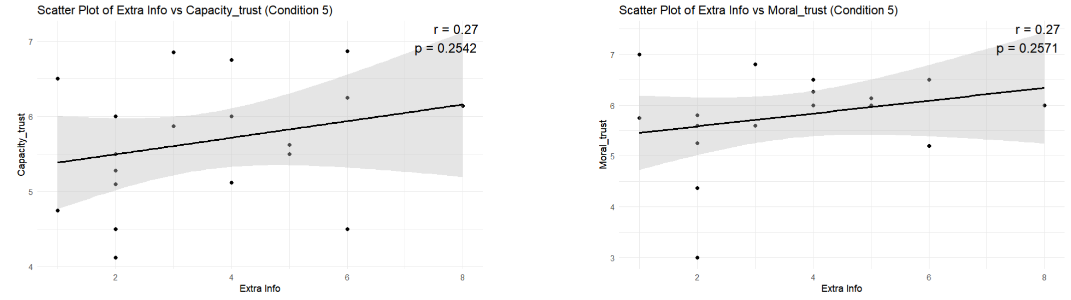


Figure 6: Correlation with capacity trust

Figure 7: Correlation with moral trust

Figure 8: Correlation between dependent variables and the number of times the on-demand explanations are required

5 Responsible Research

In order to ensure the ethical standards of the research, multiple steps were taken concerning the anonymity of the data, the reproducibility of the method, the use of the results, and the availability of the code source.

To begin with, the data collected from the participants was anonymized to protect their confidentiality. Their privacy was respected by securely storing their responses and not linking them to specific participants. Responses were collected and stored using Qualitrics and the study reported only summarized data and results obtained based on the participants' feedback.

The methods employed in conducting this research were thoroughly documented to facilitate reproducibility. This transparency aims to make the research more approachable to the readers, enabling them to replicate it and verify the results. Moreover, this ensures that the research can become a foundation for future studies. Descriptions of the study, the implemented code base, the interface the participants interact with, the questionnaires, and the data analysis have been reported. The use of established tests for the data analysis and the scaling concerning the control variables contributes to the reproducibility of the research.

The participants had no prior knowledge of the experiment, a measure taken to prevent biases. Before participating in the study, they gave their informed consent, approved by the ethics committee of our institutions (ID 4002). They were informed about the purpose of the study, the data collection methods, the procedure, and the means of respecting their privacy. The selection bias was avoided by reporting all the data gathered from participants in the analysis process.

Guided by those principles, the research ensures that the methods were transparently described and that the results are reliable and utilizable in future research.

6 Discussion

The results indicate there are no statistically significant differences when comparing the dependent variables in the baseline and the on-demand conditions. In this section, the possible reasons behind this will be further explored.

The high mean values in both conditions for the capacity trust, the moral trust, and the XAI satisfaction imply that participants generally trusted and were satisfied with the explanations of the robot, regardless of the condition type. The collected feedback for the dependent variables had a high score. The low disagreement rates in both conditions suggest that participants' agreement with the robot's decisions, can be considered consistently high. The low standard deviations indicate that the responses of the participants were consistent in both conditions.

Concerning moral trust, XAI satisfaction, and disagreement rate, a Wilcoxon test has been used on the data, their high p-values indicate a high likelihood that the differences between the conditions occurred randomly. The lack of significance could be explained by the consistently high scores for trust and satisfaction, making it harder to assess potential diversity.

Previously referenced research shows that personalized explanations have a positive impact on the trust of the human supervisor [13]. The personalized explanations mentioned in this research involve the agents who use a user model of the human to personalize their explanations. The on-demand explanations are also personalized, allowing human supervisors to ask for additional information on demand.

As has been previously mentioned, the interaction quality increases the satisfaction of the supervisor with the capabilities of the robot [18]. Therefore, the high scores on both the baseline and the on-demand condition could suggest that the baseline version has already been implemented in a manner that enhances trust and satisfaction. As a result, the quality of the interaction might not have been significantly changed by adding the possibility of requesting the visual graphs only on demand. This can be further explored by introducing an expanded variation on the types of additional explanations and varying complexity. On the other hand, the baseline explanation can be modified for future research and one without visual explanations might have resulted in statistically significant differences.

Moreover, the visual explanations might have created an information overload, diminishing the level of attention that the participants paid to the graphic explanations. Transparency of the decision-making process has been shown to enhance trust and accountability [13]. This might serve as an explanation as to why the scores were high in both conditions and there were no statistically significant differences observed. Both conditions offered textual and visual explanations, even though for the on-demand condition the visual ones were displayed only on request, so they were both using a transparent approach.

Although additional tests were conducted to assess whether the background of the users could be linked to the frequency of the requests for on-demand explanations, no significant correlations have been found. This suggests that the participant's characteristics did not impact the request for more information. Therefore, for future studies, a more diverse list of participants could help identify whether specific groups could benefit more from the additional explanations.

The results highlight that the general trust and satisfaction with the explanations of the robot was high among the participants. The participants seem to collaborate well in the context of dynamic task allocation systems. It has been previously discussed that the use of formal analysis tools [10], enables a smooth collaboration. For this research, this dynamic task allocation enables robots to adjust their tasks and explanations based on changing environmental conditions and human interventions. This can be related to the high score of capacity trust. The satisfaction with the robot's dynamic decisions and explanations can be observed by analyzing the low scores obtained when measuring the disagreement rate variable. However, this type of additional explanation does not significantly contribute to the level of trust and satisfaction.

7 Limitations and Future Work

One limitation of this research is the low diversity in the background of participants, which was particularly impactful during the data analysis process. While performing the Wilcoxon tests on the age and education control variables, significant differences in their distributions were observed between the baseline and on-demand conditions. A linear regression analysis was conducted to verify that these variables were not significant predictors of the dependent variables. However, it is recommended that future research include more participants to balance these differences better and enhance the robustness of the findings.

The same issue applies to the risk propensity control variable. Its p-value was lower than 0.05 in the linear regression analysis with the disagreement rate as the dependent variable, indicating statistical significance. Future studies should include a larger number of participants to validate these results and explore the impact of risk propensity further.

Another point to consider, based on participant feedback, is the need to improve the predictability of the robot's actions. Participants reported feeling the need to take over decision-making because they were unaware of what the robot agent was planning to do at decision points. Enhancing the system to announce the robot's intended actions in advance could address this issue.

Additionally, many participants found the timeframe to complete the tasks in the virtual environment too short, affecting the completeness variable. The timing was inaccurate, with virtual minutes passing at a different rate than real minutes. This issue could be mitigated by extending the timeframe and ensuring it aligns more closely with real-world timing.

Finally, the threshold determining whether the robot or the human supervisor makes decisions can be adjusted. As human trust in the robot increases, the threshold can be set higher, allowing the robot to make more decisions autonomously.

8 Conclusion

This research shows that having the visual explanations displayed only upon the human supervisor's request (on-demand explanations) compared to having them consistently shown (baseline explanations) did not result in a statistically significant difference in human trust and explainable AI satisfaction with the firefighting robot. The lack of correlation between the frequency of demanding additional information (3.5 on average) and the scores of the dependent variables is an indicator that the participants did not rely on visual reports to base their judgments. Overall, the participants' level of confidence in the robot's actions and decisions, as well as their satisfaction with the AI was consistently high, while their disagreement rate was regularly low.

The background of the participants did not significantly impact the dependent variables. However, to improve the balance across control variables such as age, education, and risk propensity, conducting more diverse user studies is recommended. Furthermore, participants reported that receiving additional information from the robot agent, particularly regarding its planned actions in morally sensitive situations, would be a beneficial improvement. Altogether, this research highlights the importance of the transparency and communication of the robot's decision-making process for enhancing the human-AI agent collaboration.

References

[1] J. van Diggelen, J. Bamhoorn, M. M. Peeters, W. van Staal, M. Stolk, B. van der Vecht, J. van der Waa, and J. M. Schraagen, "Pluggable social artificial intelligence for enabling human-agent teaming," *arXiv preprint arXiv:1904.04942*, 2019.

[2] J. van der Waa, J. van Diggelen, L. Cavalcante Siebert, M. Neerincx, and C. Jonker, "Allocation of moral decision-making in human-agent teams: A pattern approach," in *Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19â24, 2020, Proceedings, Part II*, pp. 203–220, Springer International Publishing, 2020.

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

[4] P. Liu, H. Yu, S. Cang, and L. Vladareanu, "Robot-assisted smart firefighting and interdisciplinary perspectives," in *2016 22nd International Conference on Automation and Computing (ICAC)*, September 2016.

[5] I. Nourbakhsh, K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion, "Human-robot teaming for search and rescue," *IEEE Pervasive Computing*, vol. 4, no. 1, pp. 72–79, 2005.

[6] Y. Liu and other authors, "Firefighting robot with deep learning and machine vision," *Neural Computing and Applications*, vol. 34, pp. 2831–2839, 2022.

[7] G. Kuznetsov, N. Kopylov, E. Sushkina, and A. Zhdanova, "Adaptation of fire-fighting systems to localization of fires in the premises: Review," *Energies*, vol. 15, no. 2, p. 522, 2022.

[8] F. Alaieri and A. Vellino, "Ethical decision making in robots: Autonomy, trust and responsibility," in *Social Robotics* (A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, eds.), (Cham), pp. 159–168, Springer International Publishing, 2016.

[9] D. H. Hagos and D. B. Rawat, “Recent advances in artificial intelligence and tactical autonomy: Current status, challenges, and perspectives,” *Sensors*, vol. 22, no. 24, 2022.

[10] K. Lerman, C. Jones, A. Galstyan, and M. J. MatariÄ, “Analysis of dynamic task allocation in multi-robot systems,” *Information Sciences Institute and Computer Science Department, University of Southern California*, 2006.

[11] S. Verdult, J. Van Diggelen, T. Haije, and I. Cocu, “Moral decision making in human-agent teams: Human control and the role of explanations,” *Frontiers in Robotics and AI*, vol. 8, p. 640647, 2021.

[12] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, “Meaningful human control and variable autonomy in human-robot teams for firefighting,” *Frontiers in Robotics and AI*, vol. 11, p. 1323980, 2024.

[13] R. S. Verhagen, M. A. Neerincx, C. Parlar, M. Vogel, and M. L. Tielman, “Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance,” in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS ’23, (Richland, SC), p. 2316â2318, International Foundation for Autonomous Agents and Multiagent Systems, 2023.

[14] M. A. Clinciu and H. F. Hastie, “A survey of explainable ai terminology,” *Edinburgh Centre for Robotics, Heriot-Watt University*, 2021.

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

[16] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, “Explainable agents and robots: Results from a systematic literature review,” in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, (Montreal, Canada), pp. 1078–1088, International Foundation for Autonomous Agents and Multiagent Systems, May 2019.

[17] S. Atakishiyev, M. Salameh, H. Babiker, and R. Goebel, “Explaining autonomous driving actions with visual question answering,” *ArXiv*, 2023.

[18] N. Churamani, P. Anton, M. BrÃŒgger, and S. Wermter, “The impact of personalisation on human-robot interaction in learning scenarios,” in *Proceedings of the 5th International Conference on Human-Agent Interaction (HAI)*, (Bielefeld, Germany), October 2017.

[19] R. Meertens and R. Lion, “Measuring an individual’s tendency to take risks: The risk propensity scale,” *Journal of Applied Social Psychology*, vol. 38, no. 6, pp. 1506–1520, 2008.

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

[21] G. Kahane, J. A. C. Everett, B. D. Earp, L. Caviola, N. S. Faber, M. J. Crockett, and J. Savulescu, "Beyond sacrificial harm: A two-dimensional model of utilitarian psychology," *Psychological Review*, vol. 125, no. 2, pp. 131–164, 2018.

[22] B. F. Malle and D. Ullman, "A multi-dimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction: Research and Applications* (C. S. Nam and J. B. Lyons, eds.), pp. 3–25, Amsterdam, Netherlands: Elsevier, 2021.

A Appendix

A.1 Correlation analysis

Here are presented more figures describing the correlation analysis.

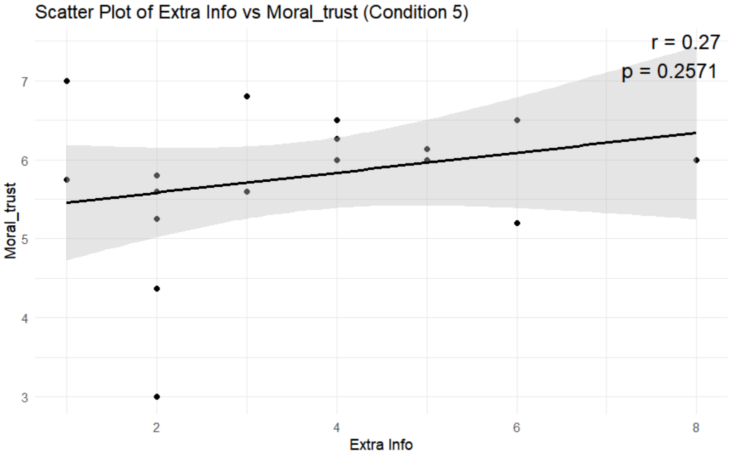


Figure 9: Correlation between extra info and moral trust

A.2 Questionnaire before interaction with platform

Here are presented the questions that were part of the questionnaire.

A.3 Questionnaire after interaction with platform

Here are presented more details about the questionnaire shown after the interaction with the platform.

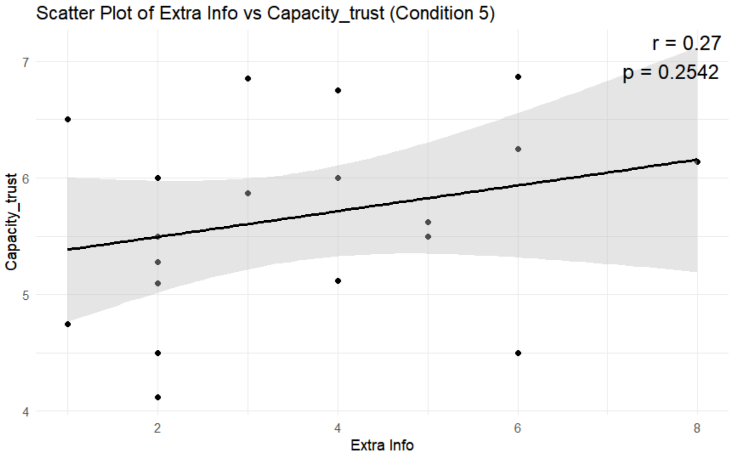


Figure 10: Correlation between extra info and capacity trust

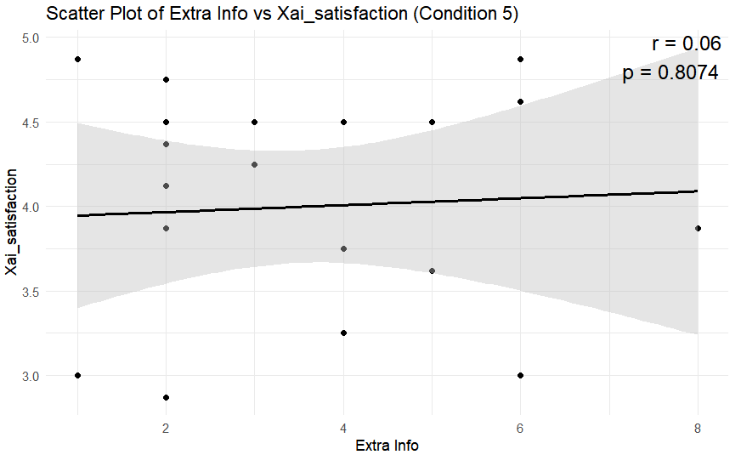


Figure 11: Correlation between extra info and XAI satisfaction

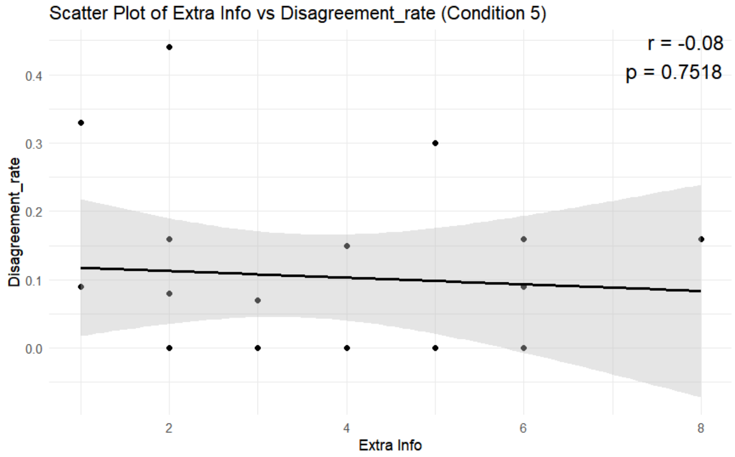


Figure 12: Correlation between extra info and disagreement rate

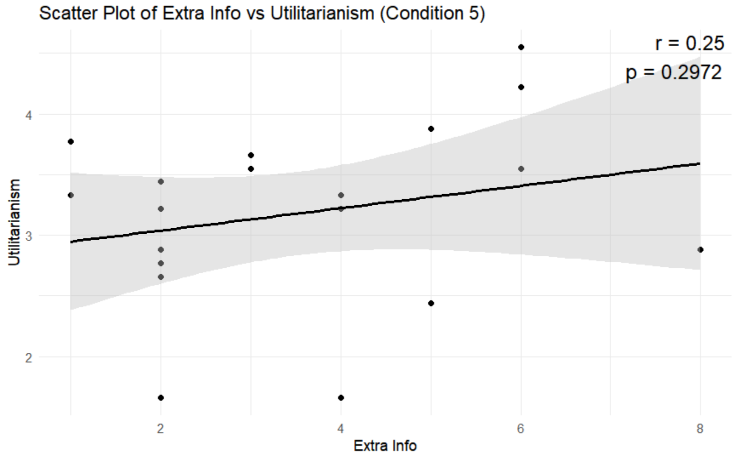


Figure 13: Correlation between extra info and utilitarianism

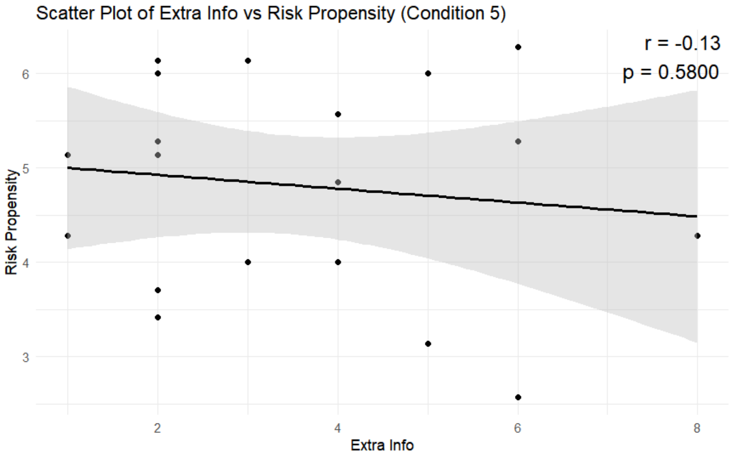


Figure 14: Correlation between extra info and risk propensity

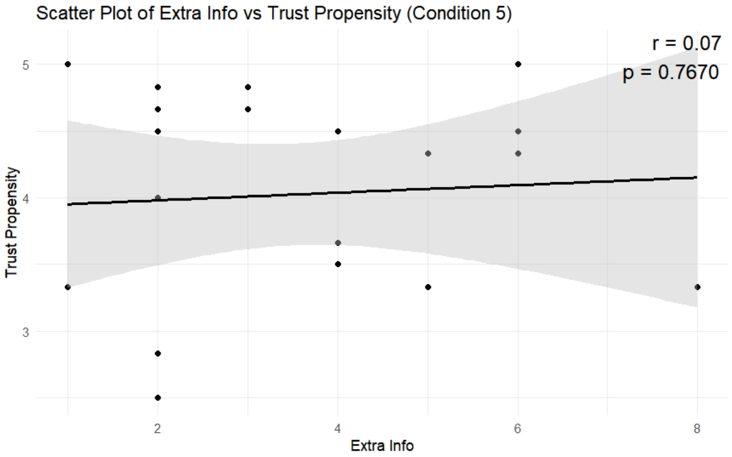
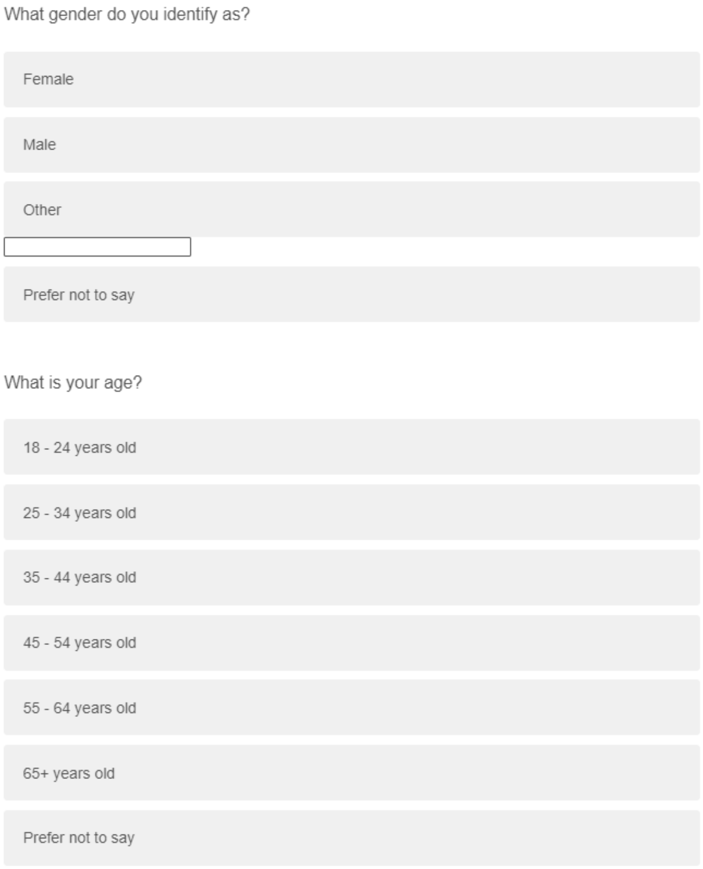


Figure 15: Correlation between extra info and trust propensity



What gender do you identify as?

Female

Male

Other

Prefer not to say

What is your age?

18 - 24 years old

25 - 34 years old

35 - 44 years old

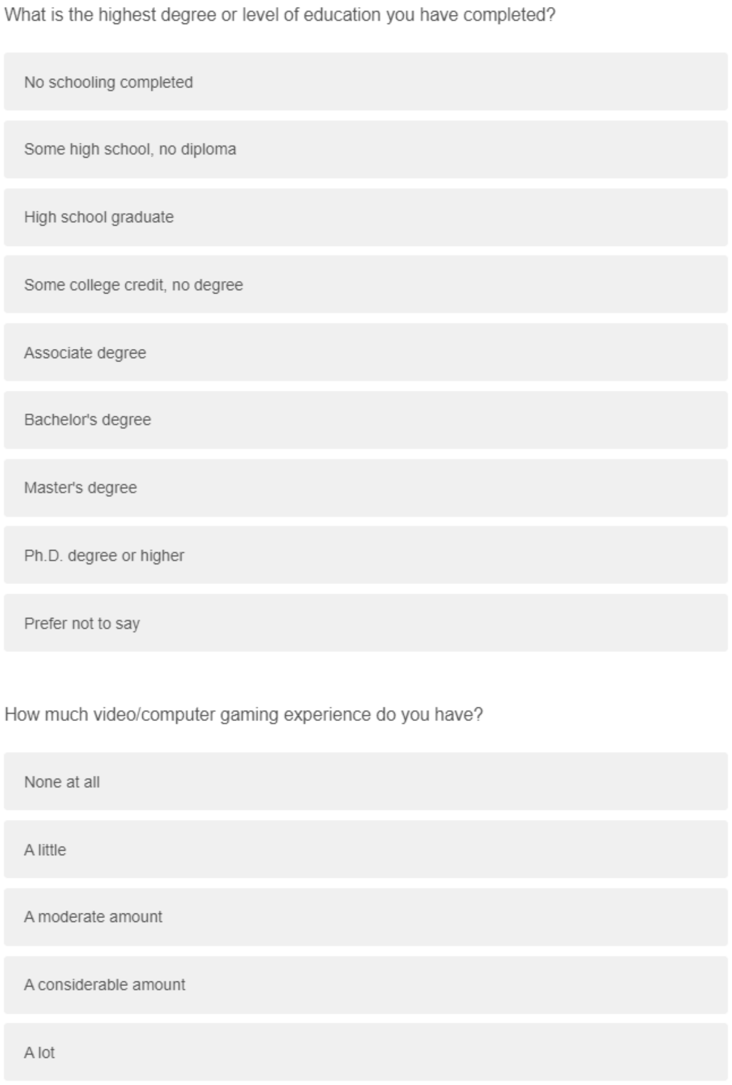
45 - 54 years old

55 - 64 years old

65+ years old

Prefer not to say

Figure 16: Gender and age



What is the highest degree or level of education you have completed?

No schooling completed

Some high school, no diploma

High school graduate

Some college credit, no degree

Associate degree

Bachelor's degree

Master's degree

Ph.D. degree or higher

Prefer not to say

How much video/computer gaming experience do you have?

None at all

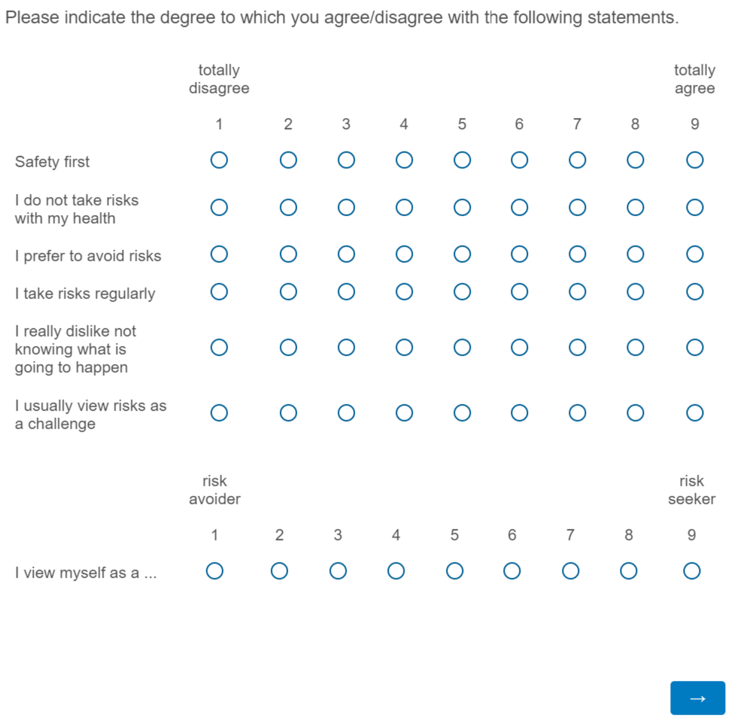
A little

A moderate amount

A considerable amount

A lot

Figure 17: Education and gaming experience

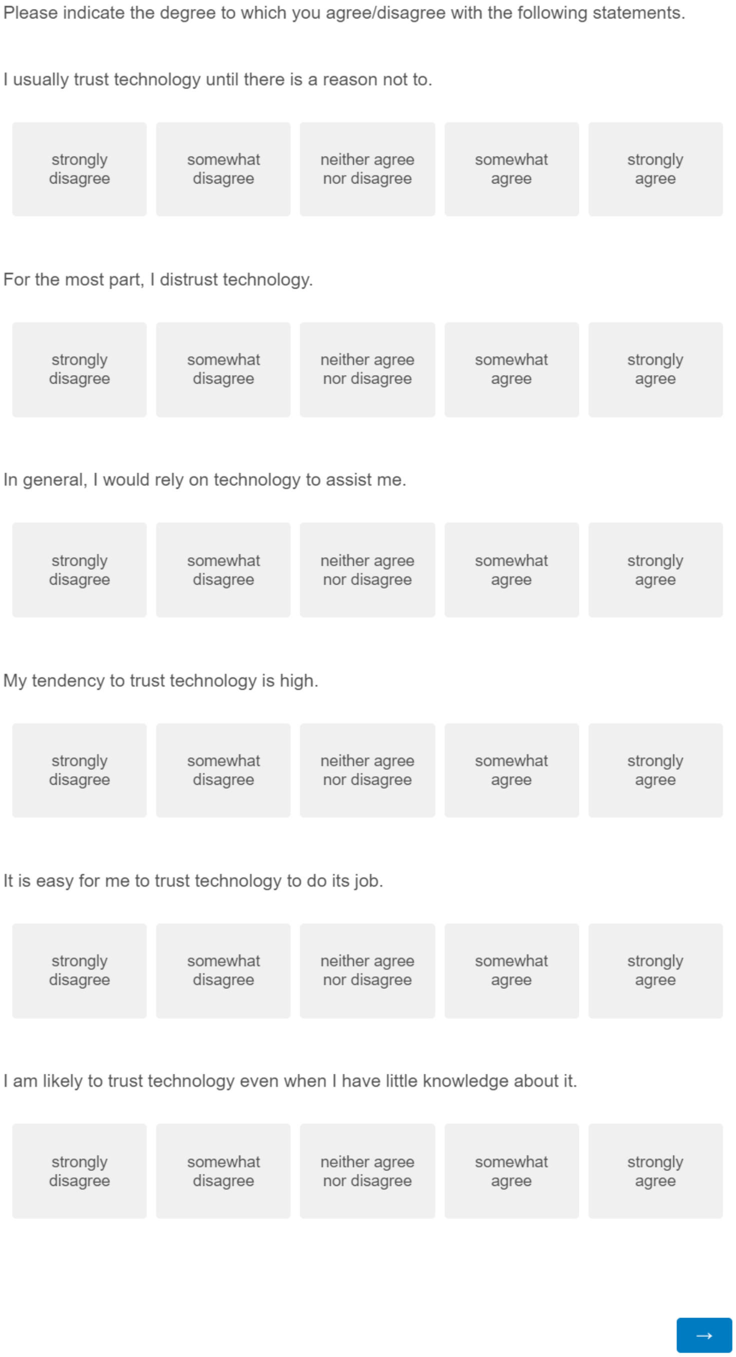


Please indicate the degree to which you agree/disagree with the following statements.

	totally disagree								totally agree
	1	2	3	4	5	6	7	8	9
Safety first	○	○	○	○	○	○	○	○	○
I do not take risks with my health	○	○	○	○	○	○	○	○	○
I prefer to avoid risks	○	○	○	○	○	○	○	○	○
I take risks regularly	○	○	○	○	○	○	○	○	○
I really dislike not knowing what is going to happen	○	○	○	○	○	○	○	○	○
I usually view risks as a challenge	○	○	○	○	○	○	○	○	○

	risk avoider								risk seeker
	1	2	3	4	5	6	7	8	9
I view myself as a ...	○	○	○	○	○	○	○	○	○

Figure 18: Risk propensity



Please indicate the degree to which you agree/disagree with the following statements.

I usually trust technology until there is a reason not to.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree

For the most part, I distrust technology.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree

In general, I would rely on technology to assist me.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree

My tendency to trust technology is high.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree

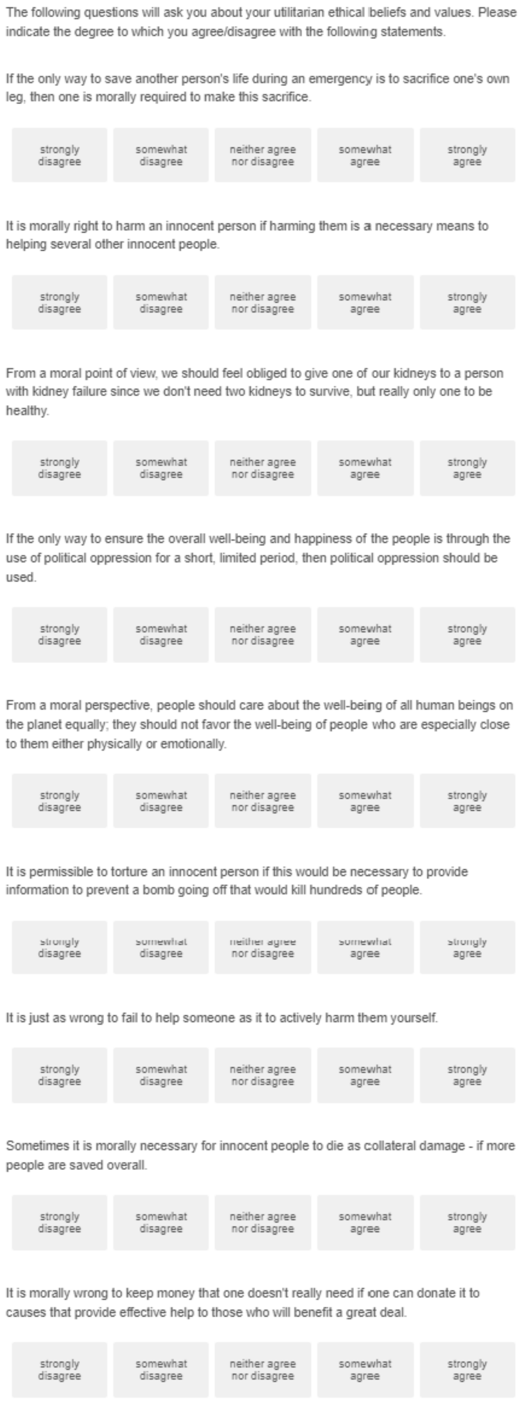
It is easy for me to trust technology to do its job.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree

I am likely to trust technology even when I have little knowledge about it.

strongly disagree	somewhat disagree	neither agree nor disagree	somewhat agree	strongly agree

Figure 19: Trust propensity



The following questions will ask you about your utilitarian ethical beliefs and values. Please indicate the degree to which you agree/disagree with the following statements.

If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.

It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.

From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.

If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.

From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either physically or emotionally.

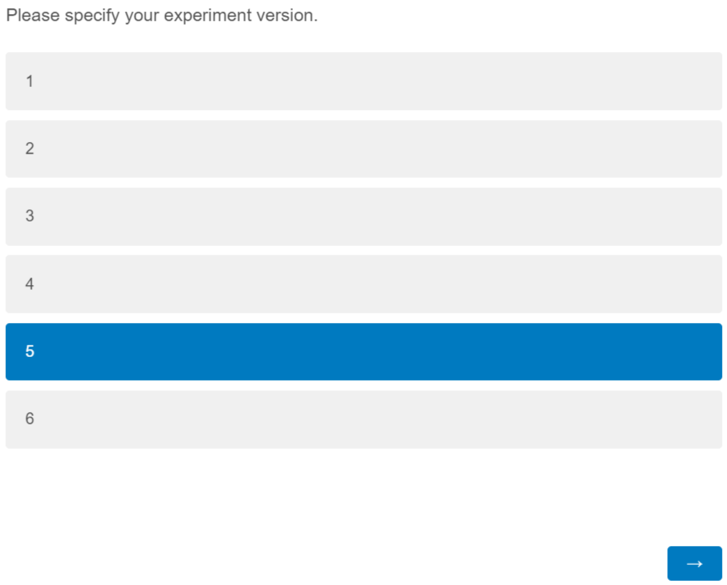
It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.

It is just as wrong to fail to help someone as it to actively harm them yourself.

Sometimes it is morally necessary for innocent people to die as collateral damage - if more people are saved overall.

It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal.

Figure 20: Utilitarianism



Please specify your experiment version.

1

2

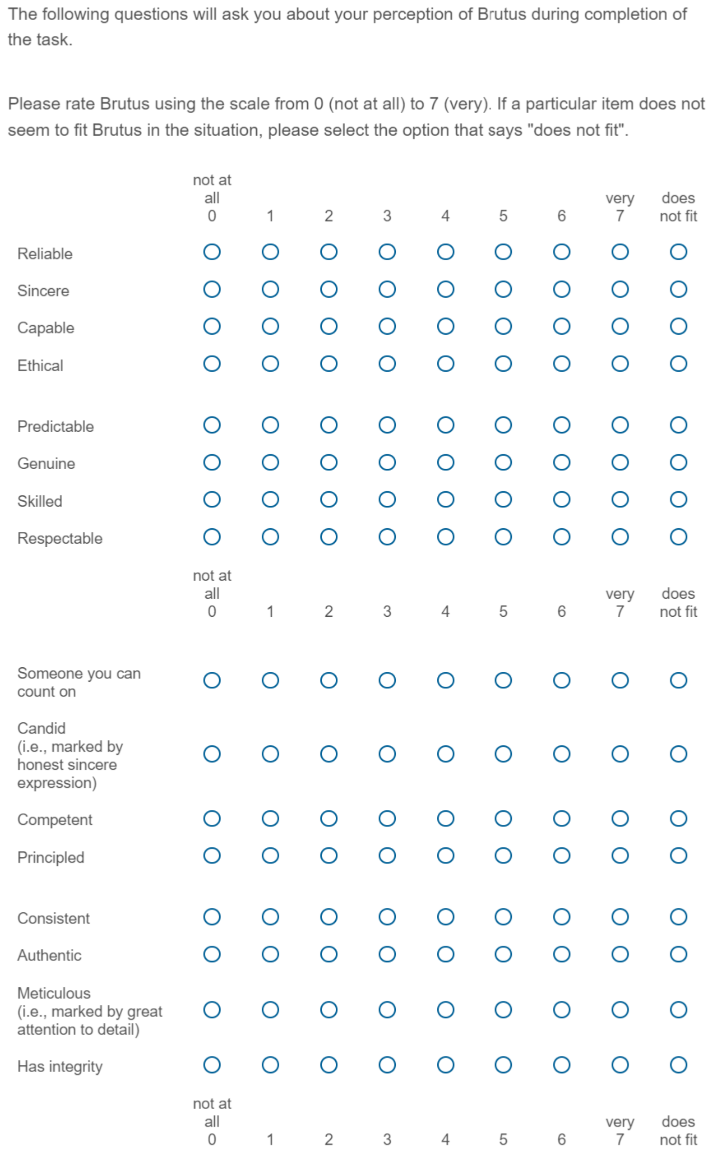
3

4

[✓ 5]

6

Figure 21: Condition of the experiment, 1 for baseline, 5 for on-demand

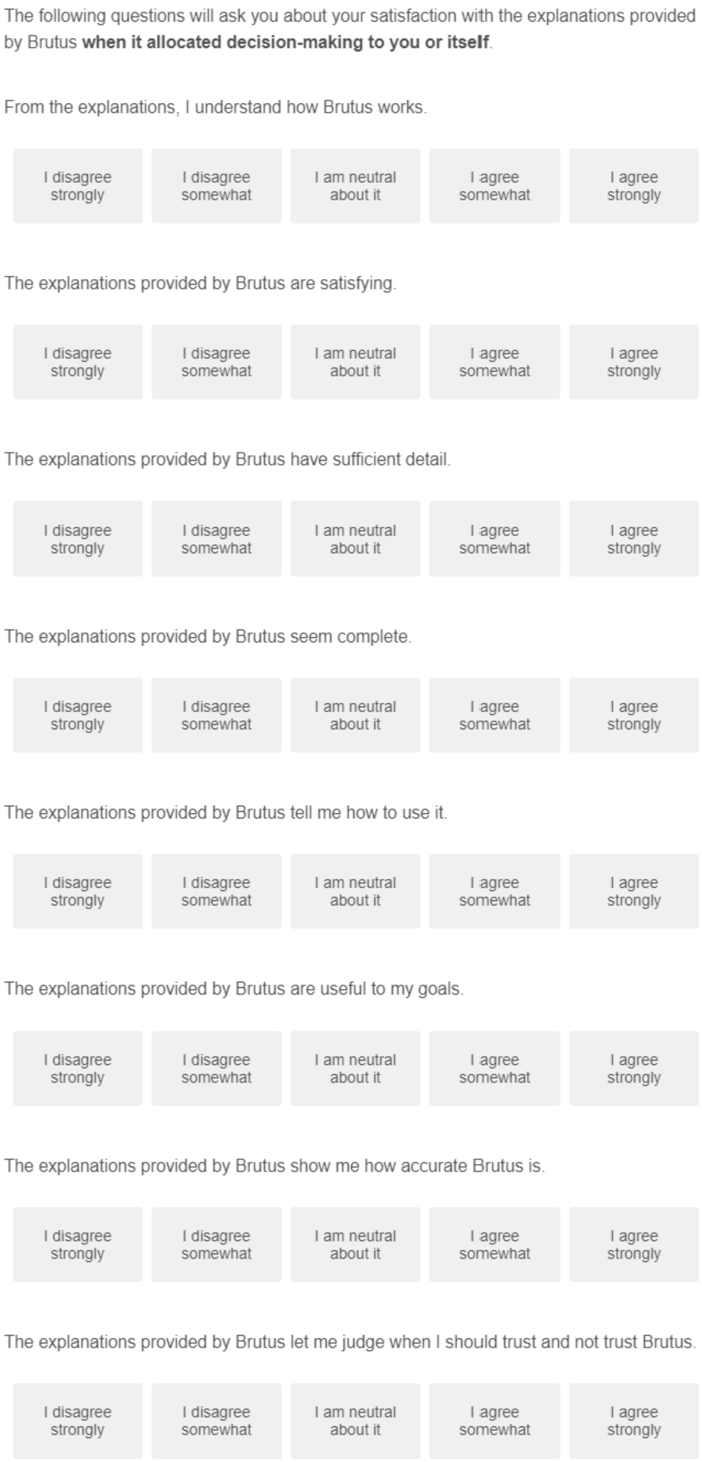


The following questions will ask you about your perception of Brutus during completion of the task.

Please rate Brutus using the scale from 0 (not at all) to 7 (very). If a particular item does not seem to fit Brutus in the situation, please select the option that says "does not fit".

	not at all 0	1	2	3	4	5	6	very 7	does not fit
Reliable	○	○	○	○	○	○	○	○	○
Sincere	○	○	○	○	○	○	○	○	○
Capable	○	○	○	○	○	○	○	○	○
Ethical	○	○	○	○	○	○	○	○	○
Predictable	○	○	○	○	○	○	○	○	○
Genuine	○	○	○	○	○	○	○	○	○
Skilled	○	○	○	○	○	○	○	○	○
Respectable	○	○	○	○	○	○	○	○	○
	not at all 0	1	2	3	4	5	6	very 7	does not fit
Someone you can count on	○	○	○	○	○	○	○	○	○
Candid (i.e., marked by honest sincere expression)	○	○	○	○	○	○	○	○	○
Competent	○	○	○	○	○	○	○	○	○
Principled	○	○	○	○	○	○	○	○	○
Consistent	○	○	○	○	○	○	○	○	○
Authentic	○	○	○	○	○	○	○	○	○
Meticulous (i.e., marked by great attention to detail)	○	○	○	○	○	○	○	○	○
Has integrity	○	○	○	○	○	○	○	○	○
	not at all 0	1	2	3	4	5	6	very 7	does not fit

Figure 22: Capacity trust and moral trust



The following questions will ask you about your satisfaction with the explanations provided by Brutus **when it allocated decision-making to you or itself**.

From the explanations, I understand how Brutus works.

The explanations provided by Brutus are satisfying.

The explanations provided by Brutus have sufficient detail.

The explanations provided by Brutus seem complete.

The explanations provided by Brutus tell me how to use it.

The explanations provided by Brutus are useful to my goals.

The explanations provided by Brutus show me how accurate Brutus is.

The explanations provided by Brutus let me judge when I should trust and not trust Brutus.

Figure 23: XAI satisfaction