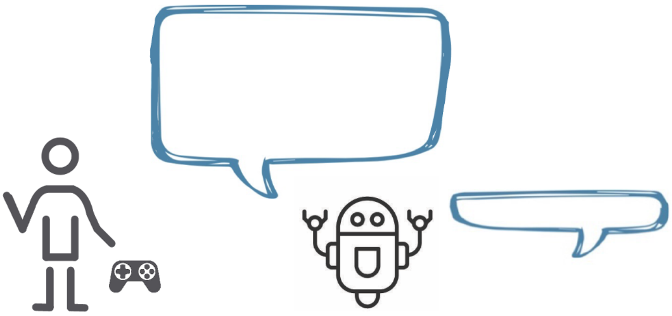


Exploring the effect of the information amount in explanation on different gaming expertise levels

by

Student Name	Jing Zhou
Student Number	5262143

Thesis advisor:	Mark Neerincx
Daily supervisor:	Myrthe Tielman
Daily co-supervisor:	Ruben Verhagen
Faculty:	Faculty of EEMCS, Delft

Style: TU Delft Report Style, with modifications by Daan Zwaneveld

Acknowledgement

Completing this thesis report means that I am now at the final step of my Master's program. During these years of studying, in addition to the accumulation of knowledge, I have also become more independent as a person. Every progress I made is not solely attributed to my efforts alone, so now I want to express my gratitude to everyone who contributed to this thesis project.

First I would like to thank my daily supervisor Myrthe Tielman, for accepting me to the research group and supervising me for a final thesis. Thanks to Ruben Verhagen for his continuous support and feedback. I am not an expert in academic writing, but Ruben and Myrthe's feedback always keeps me motivated. Thanks to our weekly meeting we can share not only our progress but also freely discuss our ideas. Thanks to Mark Neerincx for being my supervisor and also for providing valuable feedback for data analysis. Thanks to Stjepan Picek for being the external committee member, despite that my thesis topic is not closely aligned with his area.

I am thankful to all the participants and friends who volunteered to take their time and efforts to participate in this experiment, without them, I cannot complete this project. Thanks to my boyfriend for being with me every time when I was down and always listening to my nonsense with patience. Last but not least, a heartful thanks to my parents for their financial and emotional support. Though we are far away from each other during my studies, you are always my strongest support.

Jing Zhou
Delft, September 2023

Abstract

Explainable AI (XAI) has gained increasing attention from more and more researchers with an aim to improve human interaction with AI systems. In the context of human-agent teamwork (HAT), providing explainability to the agent helps to increase shared team knowledge and belief, therefore improving overall teamwork. With various backgrounds and characteristics of humans, expert video gamers are found to have better perception and cognitive ability. This study aims to study the effect of information amount in explanations on four factors: subjective workload, teamwork performance, trust, and explanation satisfaction in different expertise levels in human-agent teamwork. To investigate the research question, we designed a simulated search and rescue task, encompassing two types of explanations: the one containing less detailed information, and the other presenting more detailed information. After conducting the experiment with 42 participants, we first divided all participants into three expertise levels based on their self-reported game frequency and the mock task score in the tutorial. Then we statistically analyzed the effect of information amount and expertise levels on the subjective workload, team performance, trust, explanation satisfaction, and activity level. In conclusion, we did not find evidence that adapting the information amount in explanations to gaming expertise levels can yield an improvement in the user experience during simulated search and rescue tasks. However, subjective workload is found to have a negative effect on explanation satisfaction. For future studies, it may be worth investigating whether expert gamers require explanations with very detailed information in HAT.

Contents

1 Introduction

1.1. Background and Motivation

With Artificial Intelligence (AI) techniques becoming more and more well-developed, the advantage of AI agents' computational ability and their rational reasoning attracts researchers' attention to investigate the possibility of making humans and agents work together on the same task. Some research [64] [67] investigates how to make humans and AI agents work collaboratively in an efficient way, in which both humans and agents can make the best of their own strengths. Human-agent teamwork (HAT) requires both parties to work in a highly interactive way, which involves plenty of factors that influence teamwork performance, for example, shared knowledge in the team, both parties' trust in each other, human satisfaction with the agent, etc. However, AI agents often work as black boxes, their complex algorithms make it hard for humans to fully understand the AI agents' internal reasoning processes or trust their decisions. In turn, in a teamwork setting, this lack of comprehension or trust can lead to lower efficiency and worse teamwork performance. One of the solutions to tackle this problem is explainable AI (XAI) [23].

XAI provides explanations to its users and helps to interpret the internal reasoning or explain the behavior of AI agents. Proper explanations can have a positive effect on humans' trust in the agent, and lead to more effective teamwork [62]. For example, in a search and rescue task where humans and AI agents collaborate together, providing explanations about the agent's current situation and action helps humans understand the whole situation and plan for further actions better. To enhance interpretability and user comprehension, XAI has been increasingly deployed in decision support systems [56], healthcare and medicine [75], etc.

Given the various contexts and different styles in which explanations are provided, it is hard to provide a satisfying explanation to every user or in every context. Besides, humans' preferences can be influenced by their past experience or their own characteristics, which makes different humans likely to prefer different styles of explanations. Even though generally providing explanations helps users to understand AI systems better, providing explanations with a style that users do not prefer can decrease satisfaction with the AI system. Adaptive XAI is an approach to provide more personalized and trustworthy explanations. Adaptive XAI can be distinguished by the factor that explanations adapt to. Context-aware XAI adapts explanations to factors related to context, such as time pressure, while user-aware XAI adapts to factors related to human's personality, behavior, etc. [3]. Currently, there is a lack of research that investigates what factors of humans influence the preference for some specific explanation styles or how these factors would affect the preference. Conducting such studies can help to shape the implementation of personalized XAI systems.

When performing a task, humans are likely to behave based on their background knowledge and experience. Those with similar experience tend to learn the task more quickly with better performance, while others with novice knowledge can take more time to learn the task and try to formulate their behavior during the task. Plenty of research has reported that expert gamers tend to have better attentiveness and cognitive ability [68] [53] [65]. Better cognitive ability correlates with the ability to process more information when performing a task. Though studies such as [69] investigated whether providing more information would influence HAT, there is a lack of studies about how gaming expertise levels would affect less or more detailed explanations perceived during HAT tasks.

1.2. Research Question

Based on the background in the previous section, we propose the main research question of this study: **what is the effect of less versus more detailed agent explanations on human-agent teamwork for people with different gaming expertise?**

Related sub-questions include:

- What is the effect of less versus more detailed agent explanations on subjective workload for people with different gaming expertise?
- What is the effect of less versus more detailed agent explanations on task performance, subjective trust, and explanation satisfaction for people with different gaming expertise?
- How can we design explanations that are adapted to users' gaming expertise?

2
Theoretical Foundation

This section describes the background and theoretical foundation of this research. Based on the research questions, we aim to investigate how agent explanations adapted to human gaming expertise can influence human-agent teamwork. Hence, the related theories can be divided into two parts. The first part discusses studies related to human-agent teamwork. The second part discusses the research of explanations and explainable AI.

2.1. Human-Agent Teamwork

Though studies in AI and automation are motivated by reducing cost and exceeding the capability of humans, simply replacing humans with AI systems is not the best way to compensate for human limitations [31]. Both humans and AI agents have their advantages in different fields and circumstances. A good performance requires both skill sets and both parties of AI agents and humans [31]. Hence, researchers started to investigate the possibility of humans and agents collaborating on the same task, which is called human-agent teamwork (HAT).

Teaming can be interpreted as an instance of joint activity. Bradshaw et al. divided joint activity into three types of activity: co-allocation, cooperation, and collaboration, these three activities require different levels of interdependence from low to high [9]. Co-allocation only requires interdependence of necessary resources; while cooperation involves interdependence of activities without a shared goal; collaboration only happens when teammates have a shared objective [9]. In order to achieve successful interactions in HAT, agents need to meet human expectations of their behavior [73]. There are several factors that could influence human-agent teaming, such as interdependence, roles, and task settings.

2.1.1. Interdependence

Interdependence is the keyword for teamwork. In [34], interdependence is defined as: *"Interdependence describes the set of complementary relationships that two or more parties rely on to manage required or opportunistic dependencies in joint activity"* [34]. In the context of teamwork, when team members are dependent on each other to perform a task, they are interdependent [35].

Though plenty of studies have been devoted to increasing autonomy of the AI agents, when the agents are teaming with humans, only increasing autonomy may not necessarily increase the team's performance. Besides, designing a team is more than separating and assigning tasks to different team members. Johnson et al. highlighted the importance of managing interdependence among team members in a complex joint activity that involves humans, software agents, and robots [32]. To involve interdependence when designing a human-agent team, Johnson et al proposed a design approach called **coactive design**. Coactive design is an approach to take software agents not only as an independent tool with autonomy, but rather as an interdependent teammate to work collaboratively with [34].

To design appropriate HAT, it is necessary to understand the underlying interdependence of team members [35] [31] [33]. For example, a task can be done more efficiently by working together than two people working alone. A fully defined interdependence relationship includes both the reason and the remedy for it. The reason is what this relationship is trying to address, and the remedy is about how it is going to be addressed [34].

Soft and hard interdependence

Teamwork among humans is flexible. In some cases, an individual has the ability to do some tasks independently, but with others' help, it will significantly increase efficiency and reliability. This kind of cooperation is defined as soft interdependence in [34]. Soft interdependence happens when teammates provide mutual support to help each other, it is optional and opportunistic [35]. In contrast to soft interdependence, hard interdependence is described as a strict requirement when an action is completely dependent on the other's collaboration. For example, in a search and rescue task that requires a teaming of humans and agents, soft interdependence happens when more teammates can lift a rock faster, though a rock can be lifted by one team member with more effort. While hard interdependence can be that only humans can diagnose the injury levels of the survivors. In [34], Johnson et al. argued that to achieve great teamwork, it is crucial to include both hard and soft interdependence. However, soft interdependence is always a criterion to distinguish perfect teams from good teams [35].

2.1.2. Roles in HAT

The literature study of [67] reviewed models of human-agent teamwork and suggested that in order to make agents integral to the team, researchers need to identify what role agents play in the team. Sycara et al. proposed three roles that agents can play during HAT: agents supporting individual team members in the completion of their own tasks, agents supporting the team as a whole, and agents assuming the role of an equal team member[66]. Like in human teams, the role that agents can play in HAT is influenced by their capabilities, and different roles of the agents could influence the team settings.

2.1.3. Factors in HAT

When considering teamwork between humans and AI agents, how human teamwork models work can be a good inspiration [31]. However, due to the nature of AI agents, one of the difficulties of HAT is that agents cannot communicate with humans like humans with each other. Hence, there are several crucial factors for successful HAT, such as transparency and explainability, mutual trust, shared mental model, etc. [31] [69].

Trust

Trust is a crucial factor in teamwork. With trust, human tends to cooperate and communicate more efficiently with others. Trust is highly integrated with teamwork performance, not only because trust is the foundation of efficient teamwork, but a smooth team interaction and satisfaction also improve trust between teammates [12]. Given the context of HAT, research also found that trust is related to team performance [46]. Hence, trust is a crucial factor in evaluating HAT.

shared mental model

A shared mental model fosters mutual awareness, that team members are not only aware of their own situations but also other teammate's situations, which can promote the joint goals in the team [76]. Shared mental models in teamwork can be distinguished into shared-team-mental models (models related to team interaction) and task-based-mental models (models related to task or equipment) [45]. Studies have shown that communication in human-agent team cultivates a shared mental model, and enhances team performance [25] [76], and trust [61].

2.1.4. Search and Rescue Task

As an example of HAT, search and rescue (SAR) tasks are related to real life, and provide opportunities for humans and agents to collaborate. SAR tasks are often used by research that investigates human-agent interaction, such as [30].

In reality, SAR task is always done by a team of humans and equipment, which also provides convenience to simulate HAT. The scenario of SAR includes plenty of instant information and decision-making that could test human's ability of perception and cognition, hence requiring a relatively high expertise level. An example of decisions in SAR can be to decide who to rescue next when there are several victims to rescue at the same time.

Studies have investigated how to support SAR by Unmanned Aerial Vehicles [72] [19], decision-support systems [17] [1], etc. Hence SAR benefits from involving autonomy and AI systems into the team. Due to its nature of teaming and its trend of collaborating with AI systems, SAR is often used as a simulated scenario to investigate human-agent teamwork. In a simulated SAR task, humans and agents collaborate together to search for survivors and rescue as many survivors as possible. As the name implies, SAR task can be simply divided into two sub-tasks: search task and rescue task. During the search task, the main goal is to search for survivors blocked by obstacles or survivors without entry to a safe zone. During the rescue task, the team aims to rescue the injured survivors and move survivors to the safe zone.

2.2. Explainable AI

Currently, most AI systems' internal decision-making processes remain opaque to humans and lack transparency [23]. Explainable AI (XAI) is a method aimed at improving transparency by offering explanations that enhance the comprehensibility of AI systems for users[60]. Moreover, making the AI systems explainable and transparent is essential not only for researchers but also for those who will be affected by the system's decisions [74]. In the context of HAT, providing explanations also helps to avoid human's incomprehension in the agent [18]. Hence, the ability to explain in XAI is also crucial in HAT.

XAI can be divided into data-driven XAI and goal-driven XAI. Data-driven XAI explains black-box machine-learning algorithms, for example, explaining the results of classification or interpreting the parameters in a neural network model [78]. While goal-driven XAI refers to the explainable agents that can explain their behavior or decision to end users [3]. However, data-driven XAI is more related to machine learning or neural network models, hence it is out of the scope of this study.

Humans have a good intuition of how to explain and process explanations. Miller et al. argued that analyzing how humans explain decisions and behavior to each other is a fair starting point to investigate the methods to design XAI [49]. Hence, several architectures and frameworks of XAI are inspired by human behavior and interactions. For example, investigating how people explain something to others or how people process others' explanations helps to generate certain kinds of explanations.

Miller et al. argue that if we are to design and implement agents that can truly explain themselves, the explanations will have to be interactive and adhere to principles of communication [49].

Plenty of research and user studies investigated the effects of XAI. Although the scenarios and applications of these studies and XAI systems vary a lot, most of them show a positive effect on humans' perceptions of situations. For example, experiments of [54][71] show that algorithms providing explanations are better at influencing people than those that lack explanations. Hence, XAI helps humans to better understand both the function of AI systems and the situation.

Explainable AI can be divided into three phases: explanation generation, communication, and reception [52]. Explanation generation includes the process of how explanations are generated. Explanation communication refers to the interaction between the agent and humans with explanations. Explanation reception focuses on human's perspective, which includes user studies to evaluate the explanations. In the upcoming sections, we will review the literature related to XAI by following these three phases.

2.2.1. Explanation Generation

An ideal explanation should be a fine balance among several factors, such as a proper length, an acceptable tone, containing sufficient information, etc. Too much information in the explanations increases human's cognitive load, while too little information cannot explain the situation clearly and leads to ambiguity. Hence, explanations should have a proper length, and they can not include too much information or too little. Several studies discussed the characteristics of a proper explanation. For example, Mualla et al. claim that explanations should be parsimonious, and define parsimony as a balance between simplicity and adequacy [51]. Simplicity provides simple explanations that consider the human cognitive load, and adequacy mandates the inclusion of all pertinent information in the explanation to help the user understand the situation [51].

There are many scenarios when explanations are needed, and different scenarios and contexts require different explanation types. An explanation type defines the way information is structured and is often defined by the algorithmic approach to generate explanations [71]. Different explanation types can influence user's trust [40] or satisfaction [36]. To generate an explanation, one of the things that needs to be considered is to choose an explanation type. In this section, some frequently studied explanation types are discussed.

Contrastive explanation

Social science studies have found that when people are asking a question, they tend to search for a difference between the result and their hypothesis by asking questions in a contrastive way. When people ask a 'Why' question, they often have a hypothesis themselves, and the 'Why' question implies a question of 'Why not' [71].

Contrastive explanations can be divided into **Rule-based** and **Example-based** contrastive explanations. Rule-based explanations are "if... then..." statements, whereas example-based explanations provide historical situations similar to the current situation [71]. Example-based explanations are often used between human communications, with an example of a specific case. For instance, when humans ask why not choose to remove this obstacle in a SAR task, a rule-based contrastive explanation can be *"If I choose to remove this obstacle, I will hurt myself."* An example-based contrastive explanation can be *"This obstacle is too heavy and I recommend not removing it. Last time when I tried to remove an obstacle above my ability, I broke my arm."*

In [48], Miller et al. proposed a model of contrastive explanation and defined contrastive questions into two types: counterfactual and bi-factual questions. Research of [49] found that the behavior of explaining is always contrastive and people prefer an explanation that answers why the result is contrasted with their hypothesis.

Goal-based and belief-based explanations

Other types of explanation are **goal-based** and **belief-based explanations**. Goals can be the reason for certain behaviors. Goal-based explanations cite the agent's desire and goal of a certain action[44]. An example of a goal-based explanation can be: "I suggest bringing an umbrella because I want to help you avoid getting wet in the rain." On the other hand, belief-based explanations express the agent's belief in a certain action. For example, a belief-based explanation of why suggesting bringing an umbrella could be: "I suggest bringing an umbrella because it is cloudy outside" [36].

Confidence explanations

Confidence explanations provide the confidence or certainty level of the system's decision or its behavior. This kind of explanation is always applied in the context of decision-support systems to increase user's trust. The experiment by [4] shows that providing confidence information in context-aware mobile phones can increase user's trust in the system. Another research by [42] suggests that in the context of a mobile application, explanations containing confidence are preferred by the users. However, in the context of HAT, whether to include confidence factor in the explanations should be considered with the specific task settings.

Feature attribution

Feature attribution explanations are widely used in machine learning systems, they assign importance scores to features based on certain criteria [39]. For example, the feature attribution explanation shows the scores of how strongly each feature is relevant to the model's decision [26]. Given the topic of our study, feature attribution will not be further discussed due to its nature of explaining the feature selection in complex machine learning models.

Counterfactual explanations

Counterfactual explanations reveal what should have been different in an instance to observe a diverse outcome [22]. Given the situation when the human user failed to remove an obstacle, an example of a counterfactual explanation can be: "If you have chosen to remove this rock together with me, we would successfully remove it." Counterfactual explanations help users to predict the AI system's behavior better by requiring users to simulate two possibilities: the conjecture in which users successfully remove the obstacle with the agent, and another possibility is the reality when the user did not remove the obstacle with the agent, and failed [13].

Generating adequate and satisfactory explanations is a solid base for the next steps of explanation communication and reception. With all these types of explanations, it would be essential for researchers and XAI designers to choose or combine different types of explanations according to their study's contexts and goals.

2.2.2. Explanation Communication

In the explanation communication phase, the form and the content of the explanation are considered [52]. There are various possible forms of explanation, such as text, audio, images, videos, etc. Besides, the content of explanations can be adapted to the context or personalized by the target user [3].

Adaptive Explainable AI

These adaptive XAI systems can be divided into two parts: user-aware XAI and context-aware XAI [3]. As the name indicates, user-aware XAI focuses on adapting to different users' behavior, personality, etc. While context-aware XAI tends to take the whole environment into account. There is some overlap between these two types of XAI since some of the context-aware XAI also take the user's behavior and personality into account [3].

Current research about adaptive XAI can be divided into two approaches. The first one is to study a certain user factor and its relationship to the final performance. In this approach, studies can be done by investigating the preference of groups of people for different explanation types, such as [36], or it can be done by investigating the effect of XAI adapt to certain factors [70]. In contrast, the studies by [21] [14] used the second approach that formulated a model trained by human personality or behaviors by machine learning algorithms.

Expertise and Beginners

As we discussed previously in Section 2.2.2, there are several factors that explanations can be adapted to. One of the factors that can be adapted to is the knowledge of human users. People with different knowledge and expertise levels tend to perform tasks differently. The research of [16] designed a user study to investigate how people's domain expertise level affects their understanding of explanations for a deep learning classifier. When interacting with an unfamiliar domain, participants experienced greater difficulty accurately identifying correct and incorrect classifications. Their judgments of system correctness and explanation helpfulness also changed, and response times were longer [16]. Another study by [6] investigated how domain expertise influences the result of explanations of an AI decision-supporting system. This experiment is designed in the scenario of chess playing, in which AI agents are supporting possible moves based on the current situation. Results show that users with better domain expertise work better with the explanation, however, this experiment did not have a time limit for users to decide the next move, which indicates participants are performing the task without time pressure, and they have enough time to make decisions after reading the reasoning explanation carefully.

Research also attempts to make use of knowledge level as a factor to adapt explanations to. For example, [50] implemented a model that is adaptive to the user's current knowledge level, and results prove that adapting to the user's knowledge level improved human-agent interaction. However, this study does not focus on the generation of explanations, so explanation style is not explicitly discussed. Among these studies related to expertise level, they all conclude there is a significant difference in performance between experts and beginners.

Game Experience and Expertise

In the last section, studies based on the different domain-expertise levels are reviewed. In this section, we discuss the similarities between game experience and domain expertise, and how to further relate the research of domain expertise awareness XAI to game experience.

Several psychological studies show that people with game experience are better at processing complicated information, and video game training enhances adults' cognitive and perceptual ability [68] [53] [65]. Green and Bavelier claimed that action video games promote the broadest benefits to perceptual and attentional abilities [20].

Correlation between game experience and expertise

Frequently playing video games contributes to game experience, but does not necessarily guarantee one to be an expert in video games. The relationship between game frequency and game expertise level is not a simple positive correlation. [41] argues researchers should distinguish game experience and game expertise carefully, as playing video games frequently does not necessarily mean being an expert in playing video games.

There is also some literature discussing the relationship between cognitive and perceptual levels and game expertise. Bailey et al. argue that people with more game experience are better at reactive cognitive ability, instead of proactive cognitive ability [5]. The study by [7] argues game expertise does not guarantee better cognitive and perceptual ability as there is the possibility that a better cognitive and perceptual ability results in being an expert in video games, instead of the other way around. Cognitive and perceptual ability is more general than being an expert in video games. In other words, humans with cognitive and perceptual abilities that are above average are supposed to perform better in video games. In our case, we refer to game expertise as the ability that helps humans to be good at playing video games, no matter what the relationship between game expertise and cognitive and perceptual ability is.

2.2.3. Explanation Reception

Explanation reception reflects the explanation from the user's perspective, it concerns how well the user understands the explanation [52]. Researchers can conduct a user study to evaluate how well the designed explanations are received by users. As mentioned by [3], most research in XAI tends to use the researcher's intuitions of what is a good explanation to design explanations, hence it is highly motivated to evaluate the explanations using standardized methods. Some frequently measured factors are trust, workload, explanation satisfaction, situation awareness, etc. [47] [70] [11]. In this section, we will discuss several measurements in the explanation reception phase.

Trust

The work by [63] reviewed trust in AI, machine learning, and robotics, Siau et al. argued that trust is dynamic, and can be viewed as a combination of beliefs in benevolence, competence, integrity, and the willingness of the trustor to depend on the trustee in a risky situation [63].

There are various definitions of trust. [57] describes trust as a factor in increasing users' confidence in the system. The evaluation method of trust varies depending on the specific scenario and the researcher's preference. Different studies choose different evaluation metrics for trust. Some employed questionnaires to evaluate users' perceived trust, and some used objective metrics to measure trust in an objective approach. Human behavior can be an objective factor that can reflect human trust in the agent. As an example of using objective measurement, the experiment of [6] recorded whether the human follows the AI agent's advice or not to measure the user's trust in the system.

Workload

Workload is a factor that is highly correlated to both the task performance and user experience. An overloaded task can lead to a stressful experience, while an underloaded task can lead to boredom. Both overload and underload can result in a decrease in task efficiency and performance. Measuring workload not only helps to design a better task but also increases the efficiency and performance of the task operators. In the context of XAI, most of the workload is about mental workload, instead of physical workload. However, the mental workload of a task can be influenced by many factors and is thought to be multidimensional and multifaceted, which leads to difficulties in measuring workload definitively [10].

The study by [77] distinguished three categories to measure mental workload: measurement of task performance, subjective reports, and physiological metrics. Some examples of physiological metrics are: eye movement activity [2], electroencephalograph [38], cardiac-based assessments [29]. Though physiological metrics can measure workload from an objective perspective, the requirement of wearing the devices makes the measurements hard to implement. Besides, physiological metrics cost more than the other two categories even though studies have tried to decrease their cost. Subjective reports have been frequently used to evaluate mental workload due to their practical advantage and sensitivity supported by current data [58]. Researchers also compared three commonly used evaluation methods of subjective mental workload: the NASA Task Load Index (TLX), the Subjective Workload Assessment Technique (SWAT), and the Workload Profile (WP), results show that WP bears the highest sensitivity among these three methods, and NASA TLX shows the highest correlation with performance [58].

Explanation satisfaction

System satisfaction, sometimes named system preference is a subjective measurement, that is often evaluated by questionnaires. The study by [57] claims an explanation system needs to satisfy users' requirements and to make the system easier to use or increase user's enjoyment. Explanation satisfaction is the degree to which users feel that they understand the AI system or process being explained to them, and always measured posterior and contextualized [28]. The explanation satisfaction scale provided by [28] includes eight perspectives to evaluate the explanation from the user's perspective: understandability, satisfaction, sufficient detail, completeness, usability, accuracy, and trust levels.

Explanation satisfaction is a straightforward measurement that evaluates the user's likeness of the explanation design. Measuring explanation satisfaction is crucial to judge the explanations according to the explanation's context, and a high explanation satisfaction level helps to enhance interactions between humans and agents.

Other factors

Besides the factors we are interested in, there are some other factors that researchers found to be relevant to XAI.

Effectiveness Effectiveness is defined as the extent to help users make good decisions [57]. In other words, this is evaluated by how effective the explanation helps the overall team performance.

Efficiency: Efficiency is correlated with response time. In [57], efficiency is described as the factor to helps users make decisions faster. In the context of HAT, efficiency can be measured by idle time, the response time of accepting or rejecting the agent's decision, etc.

Situation Awareness: Research about cognitive engineering referred to situation awareness as a mental model of the current state of the environment [15]. Situation awareness is defined into three levels: perception of the elements in the environment, comprehension of the current situation, and the projection of the near future status [15].

Transparency: Transparency describes how the explanation helps to explain the internal reasoning function of the system. It is considered to contribute to increasing users' trust in the system[57]. However, transparency cannot be completely perceived by users. Hence, a subjective way of measuring transparency is the perceived transparency, it is based on users' perception of how good the explanation is explaining the internal logic [57].

Although these factors might be affected by the explanations in XAI, some of those factors are more relevant to a decision-support system instead of a human-agent teamwork context, and the other of them do not fit well in our research question, so we will not consider them as measurements in this study.

2.2.4. Explanations in Human-Agent Teamwork

After reviewing some related literature about HAT and XAI, now we will focus on some user studies that explicitly investigate the effect of explanations in HAT. The study by [55] investigated the effect of XAI on situation awareness in human-machine teaming. The results found that the benefits of XAI are not universal, instead, novices achieved a higher situation awareness with explanations, while the performance of experts decreased with explanations [55]. Studies by [11], [47] investigated the effect of transparency level on HAT, with a result of increasing the information in explanations also increased user's subjective trust in the agent and task performance in HAT. However, the study by [70] did not find increasing transparency level in explanations has an effect on trust.

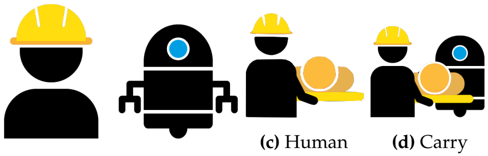
Though there are user studies investigated the effect of different transparency levels on HAT, there remains a gap in understanding whether video gaming expertise level can be a factor that affects user's experience with XAI in HAT.

3 Task design

With the knowledge of the background and related work in this field, we know that explanations generally help with human-agent teamwork, and expert gamers have better abilities to process more information at the same time. To investigate the effect of the amount of information in explanations, we first need to design two search and rescue tasks that provide explanations with either less or more information. Since we are comparing the amount of information, the only difference between these two tasks should be the amount of information in the explanation provided by the agents. Hence, in our experiment, we need to design two agents providing *less info* and *more info* explanations. The main teaming tasks that humans and agents collaborate on should be the same.

3.1. Task design

We designed a 2D collaborative game-like search and rescue task to simulate the scenario in which humans and agents work in a team. During the task, participants are supposed to collaborate with an XAI agent to rescue victims in the environment. There are two agents involved in this task: the XAI agent and the human agent, which are represented by the icons in Figure 3.1. The XAI agent is designed to work autonomously, while the human agent is controlled by the participant.



(a) Human **(b)** Agent **(c)** Human carry alone **(d)** Carry together

Figure 3.1: Icons

3.1.1. Environment

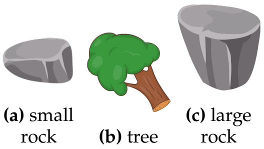
The task environment for the human's view is illustrated in Figure 3.3 and an example of god view is Figure 3.4. During the experiment, participants can only see and work in the human's view. The left side of the human's view is the map, and the right side is the chat box. During the task, the participant needs to collaborate with the agent on searching for and rescuing victims in the map. To achieve teamwork, the participant needs to monitor the chat box in order to communicate with the agent and be aware of the situation of the agent.

There are two types of rooms in this environment: normal rooms and the safe zone. Each normal room contains at least one victim, and is referred by its room index, for example, the room at the top left corner is referred as "A1". The safe zone is located at the bottom right of the environment, with an entry of dark green.

In the environment, two types of objects can be manipulated by both human and agent: obstacles and victims. Both obstacles and victims can be manipulated by the agent alone, the human alone, or both two together. The time and possibility for each action depend on the interdependence requirements.

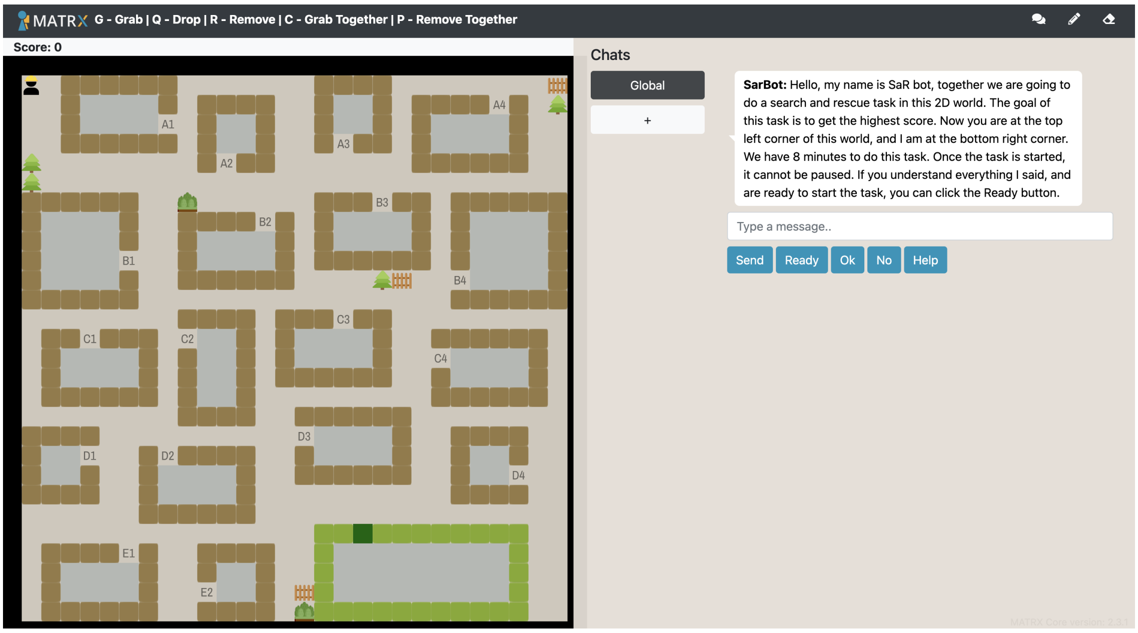
Once a victim is grabbed, it is only possible to drop him/her in the safe zone. There are 3 types of obstacles that can be removed: small rock, large rock, and tree. There are also 3 types of victims to be rescued: healthy victims, injured victims, and critically injured victims.

There are three maps involved in the whole experiment: a map for the tutorial, a map for the first task, and a map for the second task. We are using different maps to avoid participants memorizing them. The maps for the first and second tasks contain the same number of rooms and victims. There are 5 small rooms with 4 tiles, 11 medium rooms with 8 tiles, and 2 large rooms with 16 tiles. There are 25 victims to rescue in each task: 8 healthy victims for 1 point each, 12 injured victims for 3 points each, and 5 critically injured victims for 6 points each. Hence, the highest score that a participant can get is 74. Each row of the rooms contains one critically injured victim, and each room is blocked with an obstacle by default. There are no empty rooms without victims, so once a new obstacle is removed, the team can find at least one victim in that room.



(a) small rock **(b)** tree **(c)** large rock

Figure 3.2: Obstacles



SarBot: Hello, my name is SaR bot, together we are going to do a search and rescue task in this 2D world. The goal of this task is to get the highest score. Now you are at the top left corner of this world, and I am at the bottom right corner. We have 8 minutes to do this task. Once the task is started, it cannot be paused. If you understand everything I said, and are ready to start the task, you can click the Ready button.

Figure 3.3: Human view

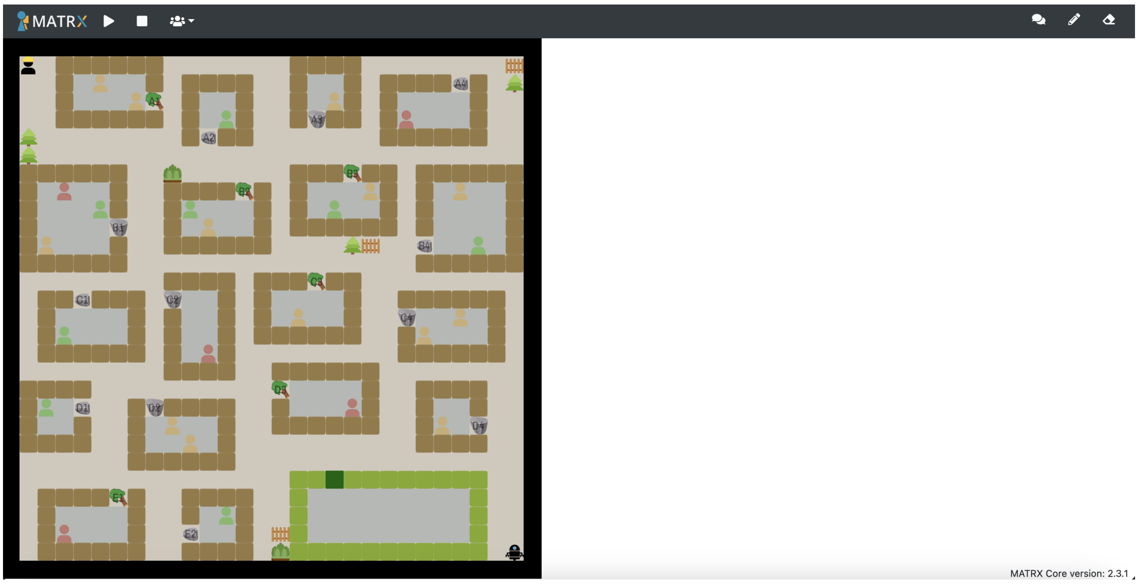


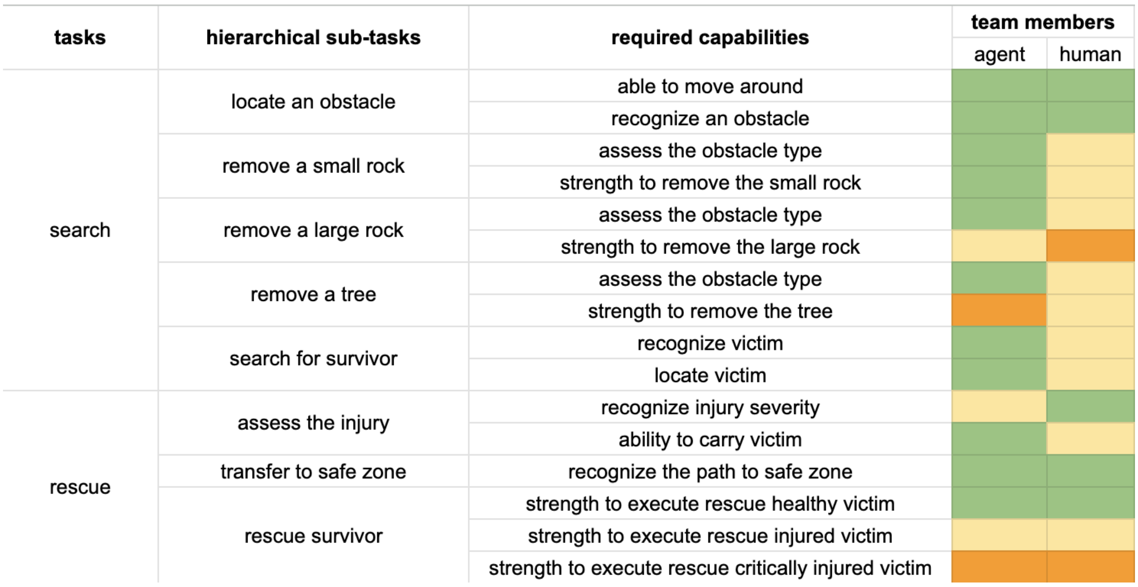
Figure 3.4: god view

3.1.2. Interdependence Analysis

To create the need for communicating via explanations in the task, we need to foster interaction between the two agents. In this section, we introduce the interdependence analysis of the task as a base to identify the information needed in the explanations. Interdependence can be divided into soft and hard interdependence. As the name implies, soft interdependence means that it is not mandatory to collaborate, but hard interdependence means teammates have to work together on that sub-task.

To simulate a real-world scenario and allow some flexibility, each of the two agents has the capability to work independently on specific tasks. Besides, to emphasize teaming, the two agents also need to collaborate on some tasks. To foster collaboration in the team as much as possible, we set different capabilities for the XAI agent and human. For soft interdependence, when actions are performed by both two agents instead of alone, the required time is reduced. For hard interdependence, the two agents have to collaborate if they want to work on the task. These capabilities are listed in Figure 3.5. The color coding in the table represents the extent of the ability to perform a certain task. Green represents the ability to do the task confidently. Yellow means it is possible to perform the task independently, but collaborating improves efficiency (i.e., soft interdependence). Orange represents this team member cannot execute the task alone (i.e. hard interdependence).

By separating the capabilities, if the agent and human want to perform a certain action, they will request each other's help. For example, when the human finds a large rock, he/she will have two options: ask the agent for help; or skip this large rock that cannot be removed alone, and search for another goal. From the perspective of the agent, if the agent encounters a situation with soft or hard interdependence and wants to request collaboration from the human, it is time for the agent to at least explain the situation.



tasks	hierarchical sub-tasks	required capabilities	team members	
			agent	human
search	locate an obstacle	able to move around	🟩	🟩
		recognize an obstacle	🟩	🟩
	remove a small rock	assess the obstacle type	🟩	🟨
		strength to remove the small rock	🟩	🟨
	remove a large rock	assess the obstacle type	🟩	🟨
		strength to remove the large rock	🟨	🟧
	remove a tree	assess the obstacle type	🟩	🟨
		strength to remove the tree	🟧	🟨
	search for survivor	recognize victim	🟩	🟨
		locate victim	🟩	🟨
rescue	assess the injury	recognize injury severity	🟨	🟩
		ability to carry victim	🟩	🟨
	transfer to safe zone	recognize the path to safe zone	🟩	🟩
	rescue survivor	strength to execute rescue healthy victim	🟩	🟩
		strength to execute rescue injured victim	🟨	🟨
		strength to execute rescue critically injured victim	🟧	🟧

Figure 3.5: Interdependence analysis table

Given the interdependence analysis, the search and rescue task can be decomposed into several sub-tasks, which help us to determine the time to provide explanations during the teamwork, and also give us a general impression of what kind of information is needed to explain during the teamwork.

3.1.3. Agent Behavior

During the task, the agent avatar runs based on the programmed agent brain. The agent is able to move around in the environment. When the agent finds a new obstacle or victim, it will send the human an explanation about the situation. The agent then waits for the human's reply for the next move. For soft interdependence, if the human agrees to collaborate, the agent will stay at its current location to wait for the human. If the human rejects to work together, the agent will start to work independently on the sub-task. For hard interdependence, if the human rejects to work together, the agent will find the next sub-task to work on. A detailed workflow for the agent can be found in Figure 3.6.

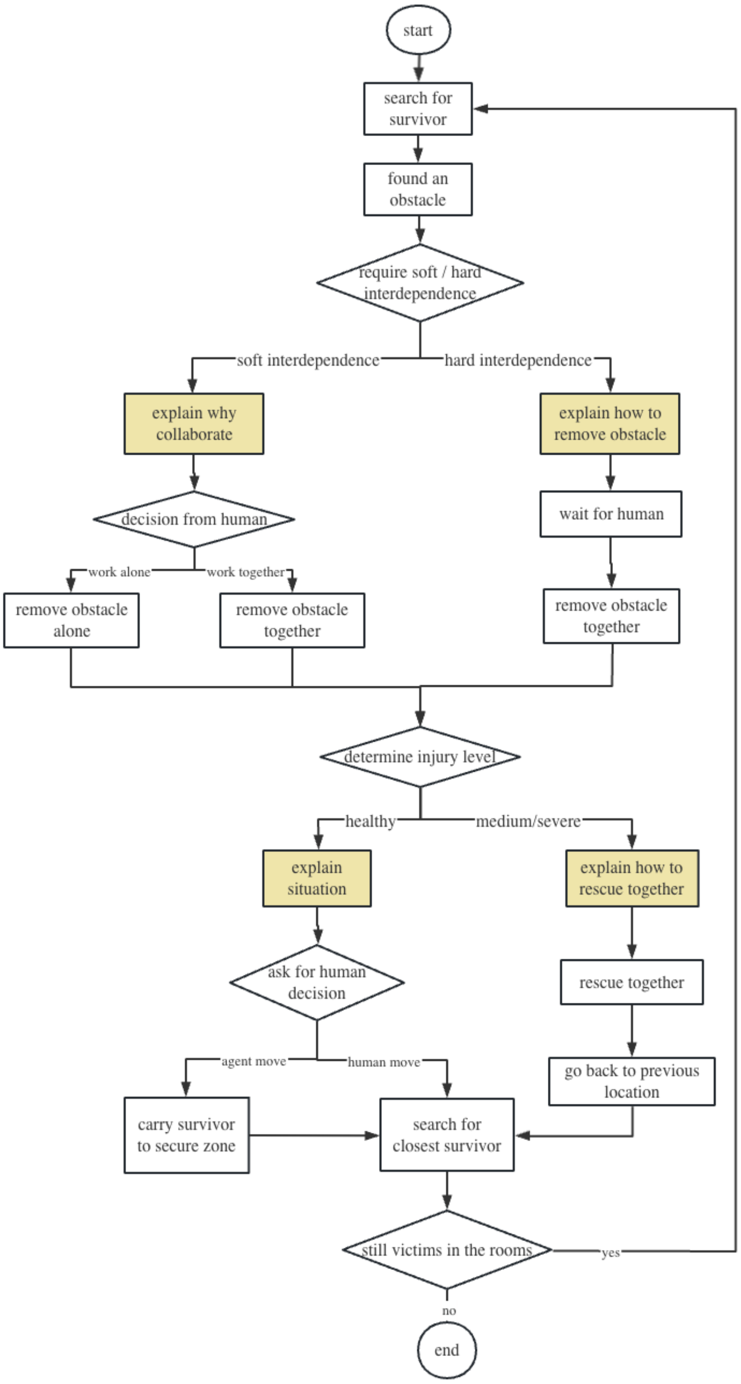


Figure 3.6: Agent working process

Based on the interdependence setting in Figure 3.5, we set the action duration for the agent's action as in Table 3.1.

Table 3.1: Agent's action duration

Action	Duration
Remove small rock alone	2.5 seconds
Remove large rock alone	15 seconds
Remove large rock or tree together with human	5 seconds
Grab a healthy victim alone	5 seconds
Grab a medium victim alone	15 seconds
Grab medium or critical victim together with human	at once

When the agent is providing an explanation for soft interdependencies, the agent first checks the distance with the human. If the distance between the human and the agent is less than 10 seconds' move and the human is not busy with some action, which saves time working together, then the agent will suggest collaborating. If the distance is too far to save time, the agent will suggest working alone. For hard interdependence, the agent will always suggest working together because it is not able to work independently.

When there are 3 minutes left for the task, the agent will send a countdown message every minute to remind the human. After the task is completed, the agent will send a message to tell the human how many victims they have rescued.

3.2. Human Behavior

The human avatar is controlled by human participants. Humans can move around with W-A-S-D keys or the four arrow keys. Besides moving, the other possible actions for humans are listed as follows:

- G: Grab a victim alone
- C: Grab a victim together with the agent
- R: Remove an obstacle alone
- P: Remove an obstacle together with the agent
- Q: Drop a victim

Except for the moving actions, the other keyboard actions are listed at the top of the task interface. During the task, the human can perform possible actions according to requirements or their own strategies. However, for each action, there are some restrictions due to the interdependence design in Figure 3.5 and common sense. The restrictions are listed below:

- G: can only grab one victim at once, victims can only be grabbed within one block away
- C: distance between the human and the agent should be less than one block; both human and agent are not grabbing victims when this action is performed
- R: obstacles can only be removed within one block away
- P: distance between the human and agent is less than one block
- Q: human has one victim when performing this action, and victims can only be dropped into the safe zone

.

When the human performs actions, it also takes a different time duration as described in Table 3.2.

Table 3.2: Human's action duration

Action	Duration
Remove small rock alone	2.5 seconds
Remove tree alone	15 seconds
Remove tree or large rock together with agent	5 seconds
Grab a healthy victim alone	1.25 seconds
Grab a medium victim alone	3.5 seconds
Grab medium or critical victim together with human	at once

In the environment, the human can always see the border of rooms and safe zone, but cannot always see the agent or the victims. The human can only perceive the agent within 3 blocks, and can only perceive obstacles or victims within 1 block. This is designed to increase the difficulty of the task, and also encourage participants to keep an eye on the explanation from the agent. Otherwise, if the human can always see the agent's location, then most of the time, it is not necessary to read the agent's explanation.

3.3. Explanation Design

To investigate the impact of explanations containing less/more information on different expertise levels, we need to design two types of explanations: explanations with less information, and ones with more information. In this study, we refer to the explanations containing less information as *less info* explanations and the ones with more information as *more info* explanations.

Although in 2.2.1 we have discussed the behavior of asking for an explanation is contrastive, it is hard for us to design a contrastive explanation because we are unlikely to know every participant's hypothesis and their 'why not' questions in this task. However, one of the main goals of explaining is to transfer knowledge to clarify others' doubts. Due to the human's limited vision in the task design, it is necessary to provide the agent's current location and situation in teamwork.

Less info explanation

According to the hypothesis, beginners with less expertise level would work better with explanations that include less information due to their relatively lower cognitive ability. *Less info* explanations should only contain the necessary information to maintain teamwork. Based on the task, when the agent provides an explanation, it needs to include the current situation and the agent's suggestion for the next action. Otherwise, the human cannot be aware of the agent's situation, and cannot know what the agent suggests to do in the next step. Explaining the situation helps humans to understand the agent's location and its new findings. Providing suggestions helps to foster teamwork by providing a decision plan. Two examples of *less info* explanations on soft and hard interdependencies are presented in Table 3.3 and Table 3.4.

Table 3.3: Less info explanation for soft interdependence

Category	Explanation
Situation	I found a large rock in room X.
Suggestion	I suggest to remove it together instead of alone (remove it by myself).

Table 3.4: Less info explanation for hard interdependence

Category	Explanation
Situation	I found a tree at room X.
Suggestion	I suggest you come here and we remove it together.

More info explanation

More info explanations are assumed to be more satisfactory to the experts. To provide an explanation that conforms to the task, we designed *more info* explanations accordingly instead of following a specific style described in Section 2.2.1. First of all, since *less info* explanations contain the necessary information for the task, *more info* explanations should contain the information in *less info* explanations. The task interdependence setting involves soft and hard interdependencies, so *more info* explanation can also be divided into two styles based on the interdependence of the scenario: explaining 'why' for soft interdependence, and explaining 'how' for hard interdependence.

An example of soft interdependence is when the agent found a large rock, which can be removed by the agent itself, but would be faster to work together with a human. Besides explaining the situation and suggestion as in *less info* explanations, the agent also explained the reason why it made a suggestion to work together or alone. In soft interdependence, the agent suggests whether working together or not depends on the distance between the human and the agent. If the human is too far away from the agent, the time saved by working together cannot compensate for the time that the human comes over, the agent will suggest to work on the task alone. Hence, when the agent explains the ability, the explanation is based on soft interdependence.

Table 3.5: More info explanation for soft interdependence

Category	Explanation
Situation	I found a large rock at room X.
Ability	I am able to remove it myself, but it would be quicker (slower) if we remove it together.
Reason	It would save (waste) x seconds if we work together.
Suggestion	I suggest to remove it together (not remove it together) instead of alone.
Following action	If you agree, I will stay here and wait for you.

For the agent, hard interdependence can be when it finds a tree that cannot be removed alone. In the scenario of hard interdependence, the agent's ability is restricted and requires other teammate's help, hence the agent's ability is one of the reasons for the agent's suggestion. The agent also provides the following action to explain 'how' to do the next step. When designing the *more info* explanations, we also tried to provide a specific method to further explain the 'how', such as providing an optimal path from the human to the agent, or reminding humans which key can perform a specific action. However, given the task design, it seems unnecessary to provide such explanations. A *more info* explanation for hard interdependence is listed in Table 3.6.

Table 3.6: More info explanation for hard interdependence

Category	Explanation
Situation	I found a tree blocking room X.
Ability/Reason	I am not able to remove it myself.
Suggestion	I suggest you come here and we remove it together.
Following action	If you agree, I will stay here and wait for you. If you disagree, I will keep searching for the next obstacle.

To summarize our explanation design, *less info* explanations provide the necessary information to maintain the teamwork between the human and the agent, while *more info* explanations provide more reasoning information to support the agent's decision.

4

Evaluation Methodology

After we have designed tasks and explanations containing less/more information in the last chapter, an experiment containing game-like search and rescue tasks is designed and conducted. In this section, the methodology of the experiment and its evaluation are described.

4.1. Hypothesis

The independent variables are the ones being controlled during the experiment. Given our research question, there are two independent variables in this study: the first one is the level of expertise; and the second one is the amount of information in the explanation, specifically in our study, less or more information. The dependent variables are those that are assumed to be affected by manipulating the independent variables, hence three dependent variables are objective task performance, subjective trust, and subjective explanation satisfaction. In this study, we consider subjective workload as a mediating variable, which can be affected by the two independent variables, and may also have an impact on dependent variables.

The conceptual model in Figure 4.1 shows the relationship between the variables.

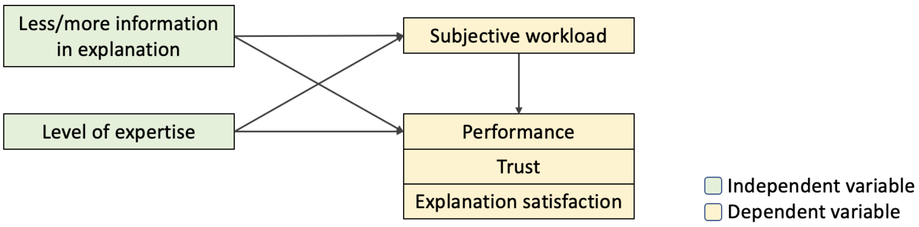


Figure 4.1: Conceptual model

Based on our research questions and background knowledge, three hypotheses related to the conceptual model are listed in Table 4.1, Table 4.2, and Table 4.3.

Table 4.1: Hypothesis1

Null hypothesis (H0)	Alternative hypothesis (H1)
Explanations with less or more information do not make a difference for people with different gaming expertise levels.	Providing explanations with more information can increase performance, trust, and explanation satisfaction for expert gamers, while explanations containing less information increase performance, trust, and explanation satisfaction for beginners.

Table 4.2: Hypothesis2

Null hypothesis (H0)	Alternative hypothesis (H1)
The amount of information in an explanation does not influence subjective perceived workload.	The amount of information in an explanation has an effect on the subjective perceived workload.

Table 4.3: Hypothesis3

Null hypothesis (H0)	Alternative hypothesis (H1)
Subjective workload does not influence the HAT performance	Subjective workload has an effect on the HAT performance.

4.2. Pilot

To pre-investigate our experiment design, a pilot study was implemented before the formal experiment. Five participants were involved in the pilot study and provided feedback on the game design and the open-question design. Two of these five participants were self-reported video game experts, while the other three did not often play video games.

The questionnaire after the tutorial is added after the pilot study. In the pilot, one of the participants mentioned that the questionnaire after the first task seemed to be a hint to the second task. Hence, we added a simplified questionnaire after the tutorial, in order to make the results of the two questionnaires after the two tasks equally biased. Since the questionnaire after the tutorial is not relevant to the research questions, these answers are not analyzed in this study.

The pilot study showed some flaws in the game design and open-question design. Four out of five participants mentioned they would like to receive a response from the agent. Once the human sends the agent a message, the agent should send the human a simple reply. Otherwise, participants would not know whether the agent received their message or not. During the pilot study, there was no ceiling effect in the self-reported expert group and the beginner group.

4.3. Participants

Participants of this study were recruited by personal connection and online advertisement. Since this study requires participants to interact with the agent in English, all participants are required to have a similar comprehension level of English. Besides, to make the results unbiased as much as possible, we did not specify game expertise or require specific game expertise levels during recruiting. The following section describes the demographic data of the participants.

4.3.1. Demographic data

In the experiment, every participant is doing the same thing, the only difference is the order of the two tasks to mitigate the learning effect. So in the demographic data, we divide them into two groups by the order of performing the two tasks. *Group1* first plays with the agent providing *less info* explanations, and *group2* first plays with the agent that provides *more info* explanations. To make sure there is no significant effect of the order of tasks on the task score, the division between *group1* and *group2* will only be used to check whether there is a significant difference between *less info* task and *more info* task on the task scores. The number of participants in each group is 21.

Age range

The bar plot of the age range between the two groups is visualized in Figure 4.2. The age range is divided into 4 groups: 18-21, 22-25, 26-29, 30 or above. After the experiment with all participants, there is no participant falls into the age range of 18-21. Thus, we discarded this group in the visualization.

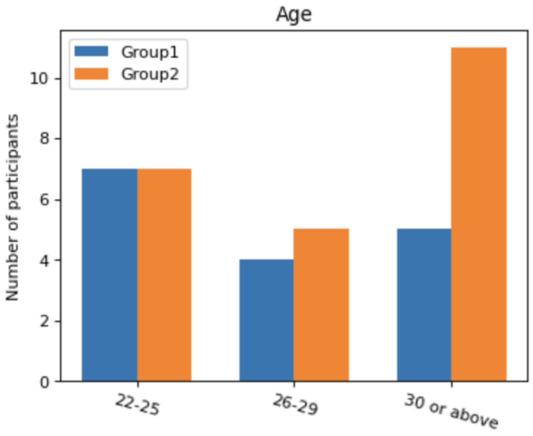


Figure 4.2: Age range distribution of two groups

To check whether there is a significant difference between the age distribution of the two groups, we performed a Mann-Whitney U test on the age distribution of the two groups. Based on the results (statistic=2.0, p-value=0.369), there is no significant difference in age range.

Gender

The bar plot of gender distribution between two groups is visualized in Figure 4.3. In the questionnaire, the gender of the participants is divided into 4 groups: 'male', 'female', 'non-binary', and 'prefer not to say'. None of the participants reported gender as non-binary. Thus, we discarded the 'non-binary' group in the visualization.

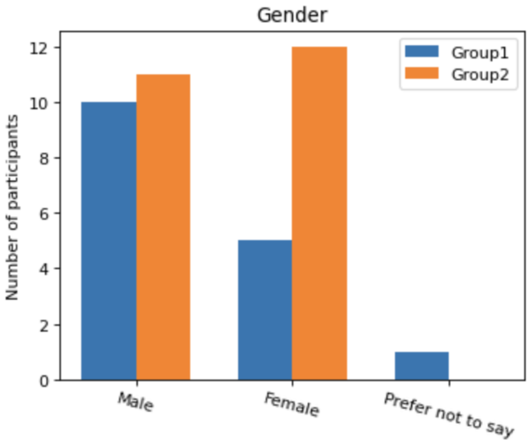


Figure 4.3: Gender distribution of two groups

We performed a Mann-Whitney U test on the gender distribution of the two groups. Based on the results (statistic=3.0, p-value=0.7), there is no significant difference in the gender of the two groups.

Educational level

The educational level is classified into four groups: 'High school or equivalent', 'Bachelor's or equivalent', 'Master's or equivalent', 'PhD or equivalent'. According to the results, none of the participants' educational level is 'High school or equivalent'. Thus in the visualization of Figure 4.4, we discarded the 'High school or equivalent group.

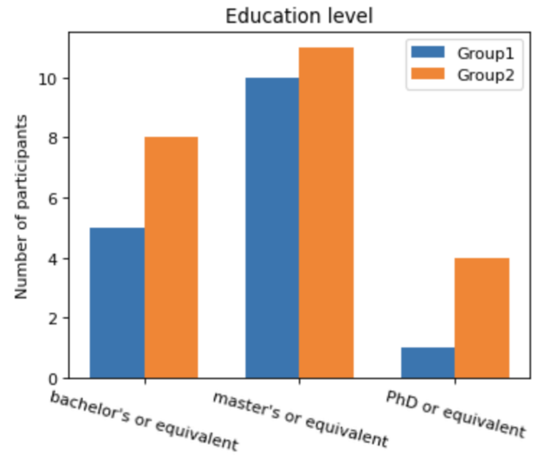


Figure 4.4: Educational level distribution of two groups

We performed a Mann-Whitney U test on the educational level of the two groups. Based on the results (statistic=3.0, p-value=0.7), there is no significant difference in the educational level of the two groups.

Game frequency

In the questionnaire, the answer for game frequency is designed as a slider, with 5 markers: never, several times a year, several times a month, several times a week, and every day. These five frequencies are represented by 0, 25, 50, 75, and 100 respectively. When we processed the data, we classified the values into four groups by their values. Frequency values falling into 0-25 are classified as ’less than several times a month’; ’26-50’ is ’several times a month’; ’51-75’ is ’several times a week’; and ’76-100’ is ’almost every day’. The self-reported game frequency of the two groups is visualized in Figure 4.5.

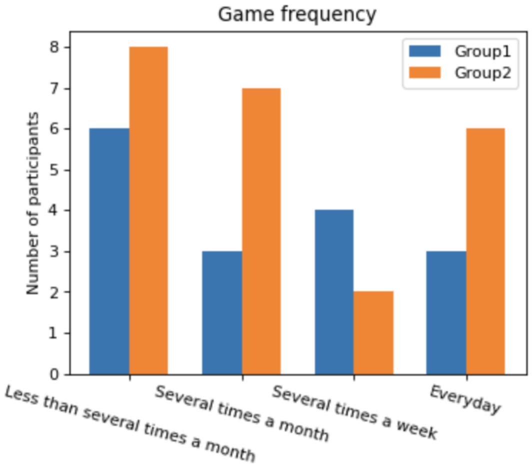


Figure 4.5: Game frequency distribution of two groups

We performed a Mann-Whitney U test on the self-reported game frequency of the two groups. Based on the results (statistic=4.5, p-value=0.381), there is no significant difference in the self-reported game frequency of the two groups.

4.4. Measurements

Since we are interested in the user experience of the explanations containing less or more information, we used subjective measurement to assess the dependent variables (workload, trust, explanation satisfaction). However, only relying on subjective measurement itself may not fully reflect what happened during the experiment. To address this problem, except for the subjective measurements, we collected some data as objective measurements during the experiment as well. In the following sections, we will first introduce the subjective measurements and questionnaires used in this study, and then explain the objective measurements.

4.4.1. Subjective Measurements

We measured subjective workload, trust, and explanation satisfaction separately with three different questionnaires. The details of the three questionnaires will be explained in the following sections. The questionnaire is attached to Appendix C.

Subjective workload

Workload can be a mediating variable between the amount of information and team performance, trust, and user satisfaction. To investigate the effect of workload, we measured subjective workload by the questionnaire from the NASA workload TLX [27]. In the TLX, subjectively perceived workload was measured in six dimensions: mental workload, physical workload, temporal demands, frustration level, effort, and performance [27]. Each dimension is measured by one question, with six questions in total. The full NASA workload TLX contains a pairwise comparison to weigh the six dimensions by participants' subjective importance. However, there is a study that supports the TLX without a pairwise comparison might increase experimental validity more than the full one. To avoid a burdensome questionnaire, in this experiment, we used the "Raw TLX" instead of the full one.

Explanation Satisfaction

Explanation satisfaction is a posterior judgment that is evaluated by system users to measure the degree to which users feel they sufficiently understand the AI system [43]. Hence, comparing explanation satisfaction to different explanations can be an explicit method to evaluate users' preferences for different explanations. To investigate the explanation from the user's perspective, we used the questionnaire from [43] to measure the explanation satisfaction. There are seven key attributes of explanation satisfaction in the questionnaire provided by [43]: understandability, the user's feeling of satisfaction, the sufficiency of detail, completeness, usefulness, trustworthiness, and accuracy. Each key attribute is measured by one question in the questionnaire. In this experiment, the explanation is the messages sent by the agent, so in the questionnaire, we rephrase "explanation" to "the messages from the bot". Participants' responses were collected by a 5-point Likert scale.

Trust

As we mentioned previously, trust can have an effect on the result of teamwork as well. In this study, we are only measuring subjective trust, and we used questions from [43] to measure users' trust in the agent. The whole questionnaire contains eight questions. The first five questions ask participants directly whether they are confident in the XAI system, and whether the XAI system is predictable, reliable, efficient, and believable [43]. The last three questions are adopted from other trust scales, and ask whether participants are wary of the XAI system, whether participants think the system can perform better than novice humans, and to what extent would participants like to use the system for decision-making. To adapt the questionnaire to our context, "XAI system" was rephrased to "SAR bot". Participants' responses were collected by a 5-point Likert scale. The sixth question measures trust in a reverted method, and the higher the score it gets, the lower trust participants perceived. Hence, before analyzing the results, the answers to the sixth question will be reversed.

Open questions

To collect feedback and other subjective opinions, we also provided 6 open questions regarding participants' suggestions for the explanation and the task. The first open question asks the participants' subjective preference for the two agents. If there is a preference for one of the agents, participants need to provide a reason for that. The second question collects participants' subjective strategies in the tasks. The third and fourth question asks participants' subjective perceived idea of how many explanations they read after the countdown message and in the overall task. The fifth and sixth questions ask for participant's feedback and suggestions for the explanation and task design. The last two questions are not mandatory.

4.4.2. Objective Measurements

The objective measurements were measured both explicitly and implicitly during the experiment. When participants were performing the tasks, the team score was displayed at the top left corner of the interface, so they could always check their current team score. Besides, the intermediate data of whether humans sent messages or performed any actions in every tick was also logged implicitly during the task.

Team Performance

Since this study mainly focuses on how well participants perform in the context of human-agent teaming, the individual performance of the human or the agent was not considered. The overall performance was measured by the final task score of the team. The minimal score that a team can get is 0, which means did not successfully rescue any of the victims, and the maximal team score is 74, implying the team has successfully rescued all of the participants. This also helped to foster the collaboration of the whole team, instead of team members competing with each other for a better individual score.

Activity Level

Besides the questionnaire, we also collected participants' log data while performing the task. The time when participants sent a message to the agent and the time when participants performed certain actions are collected by log data during the experiment. Due to the soft interdependence of the task, when participants work on the task, they can choose their own strategy to some extent.

The **response time** is the time when the agent waits for human's response to the last message. To avoid the influence of duplicate messages from humans, we only count the messages from humans after the agent already sent a message. Another variable that can reflect the activity level during the task is **idle time**. We measured idle time as the time when participants did not take actions of moving around, sending messages to the agent, and waiting for the current action to complete. Hence idle time can be interpreted as a measurement of how participants were focused on the task. The smaller the idle time is, participants tend to be more focused on the situation in the task. The last variable that is related to the activity level is **total moves**, which calculated how many tiles the human walked through during the task.

4.5. Hardware and Software

There are several materials used in task design and the questionnaire. The task environment and the agent are designed by MATRX, a library for human-agent teamwork based on Python [1]. MATRX provides several basic features for HAT design. The questionnaire is designed in Qualtrics [2], an online questionnaire designing tool. The task is run on a MacBook Pro. Both the task and the questionnaire are displayed via the Google Chrome browser.

4.6. Experiment Procedure

The procedure of this experiment is illustrated in Figure 4.6. The whole experiment takes around 30-40 minutes, depending on the time that participants spent on the tutorial and questionnaire. After the participant arrives, the instructor requests the participant to read and sign the consent form in Appendix B. If the participant has any questions related to the experiment procedure, the instructor will answer them, and then the experiment starts.

The participant will first be requested to fill in the demographic data and finish the tutorial. During the tutorial, the agent introduces the rules and abilities by sending messages to humans. The texts displayed in the tutorial are presented in Appendix A. After the introduction of all possible actions, there will be a 3-minute mock task to simulate the actual task later. The aim of this mock test is two-fold. Firstly, participants can get used to the workflow of the agent, and try to learn about the environment as much as possible. This helps to keep the surge of learning effect within the tutorial. Secondly, this mock test is an implicit measurement of gaming expertise level. Once the tutorial is finished, participants will answer a short questionnaire regarding the SAR bot in the tutorial.

After the tutorial, participants will have a basic understanding of the rules of the tasks, they will finish two game-like search and rescue tasks with two XAI agents: less info agent, and more info agent. Every participant will finish two tasks in the experiment, the order of these two tasks is randomized to

[1] MATRX: https://matrx-software.com/

[2] Qualtrics: https://www.qualtrics.com/

avoid the learning effect and the influence of first impressions. Each of these two tasks is followed by a questionnaire regarding the agent and the workload of the specific task. Once the two tasks are finished, participants will fill in six open questions about the overall experience and feedback. After the feedback session, the experiment ended.

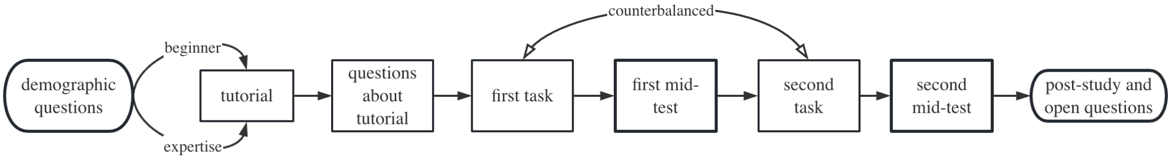


Figure 4.6: Experiment procedure

4.7. Analysis

After we collected all the results, we used Python and R to analyze the data. The statistical analysis was performed by Python and statistical libraries: $scipy$, $numpy$, $pandas$, $math$, csv. The visualization of boxplots was implemented by Python's library $Matplotlib$ and $ggplot2$ from R. The two-way mixed ANOVA was performed by $anova\_test()$ function from $rstatix$ package, and the nonparametric mixed ANOVA was performed by $nparLD$ package. The regression analysis was implemented via R and libraries lm, $gvlma$.

5 Results

Given the data collected from 42 participants, in this chapter, we will focus on presenting and analyzing the results of the experiment.

5.1. Order effect and learning effect

Our experiment was designed in a counterbalanced method, in which all participants need to complete both *less info* and *more info* tasks, but the order of the two tasks is randomized. After we collected data from all participants, we first needed to make sure the order of the two tasks did not make a difference in the task performance, and that participants' learning effect during the two tasks was not significant.

Order effect
To test whether there is a difference between the order of the two tasks, we first tested the distribution of the two tasks in Group 1(first completed *less info* task) and Group 2 (first completed *more info* task). Both of the scores in the two groups are normally distributed. Then we did two t-tests to test the difference in *less info* and *more info* task scores. There are no significant differences between group 1 and group 2 in both *less info* (statistic=0.547, p-value=0.586) and *more info* (statistic=0.889, p-value=0.378) tasks, so we can infer the order of the two tasks has no significant effect on the task scores.

Learning effect
To check whether there is a learning effect in the two tasks, we also compared participants' task scores in their first task and their second task. Since the task scores from two tasks are measured from the same population, and both the two task scores did not violate normal distribution, we performed a paired samples t-test to test the difference. The results did not show a significant difference between the first task and the second task scores (statistic=1.569, p-value=0.124). Hence, we did not observe a significant learning effect between the first task and the second task performed by all participants.

5.2. Division of experts and beginners

To answer our main research question, we first need to divide all participants into different expertise levels. There are several metrics that can be used to divide expertise levels, the most explicit one is the tutorial's task score. Since the mock task in the tutorial is the first task that participants performed without any training, we first checked whether the mock task score itself can be used to divide participants. However, the mock task score is either not normally distributed (Shapiro-Wilk p-value=0.013) or not initially clustered into several groups, so it is not ideal to divide participants only by mock task score.

Though we collected participants' self-reported game frequency as demographic data, it may be affected by participants' bias and hence does not always reflect objective expertise. The Pearson's correlation score between self-reported game frequency and the total score of three tasks is 0.203. We also visualized a scatter plot Figure 5.1 of self-reported game frequency and the total score from three tasks. Since we believe expertise should not only be focused on self-reported data but also on objective performance, we choose not to use self-reported game frequency itself to divide expertise levels. Task performance can be an objective factor in distinguishing beginners and experts, but in this research, task performance is one of the dependent variables that we are going to observe, hence cannot mix task performance with expertise levels.

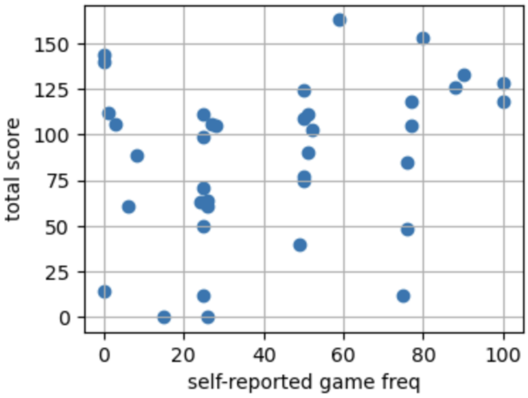


Figure 5.1: Self-reported game frequency and total score

The ones who play video games more often tend to obtain more knowledge of video games, while those who play video games less often may not be experts in video games. According to the data we have collected, we decided to use both self-reported game frequency and tutorial task score to divide expertise levels. The self-reported game frequency can be seen as a subjective factor that reflects the participant's experience in video games, and the tutorial score is an objective factor that represents how well the participant performed the task right after we explained the rules of the task.

After we plotted a scatter plot of participants' game frequency and their tutorial score, we first tried to use K-means clustering to divide all participants into two or three expertise levels. However, the nature of the data we collected cannot be clustered into several sensible clusters, so we exclude this method. After considering some possible division methods, we chose to simply divide participants using the mean of game frequency and the tutorial score. We first calculated the average self-reported game frequency and the average tutorial score from all participants. To make the division more reliable, we applied a 10% interval around the mean frequency and scores instead of using the mean value for a rigid division.

The table in 5.1 listed the details of how participants were divided into three expertise levels. The division can be visualized in Figure 5.2. The general idea of this division is that beginners are the ones who are neither not frequent gamers nor perform below average in the tutorial. The experts are the ones who are both frequent gamers and also performed better than average in the tutorial. The only exception is that there are two participants who exhibited excellent performance in the tutorial but did not classify themselves as frequent gamers. Due to their exceptional tutorial scores, we also categorized these two participants as experts. The remaining participants were counted as intermediate level. After this division, we get 14 beginners, 16 intermediate gamers, and 12 experts.

Table 5.1: Details of participants' division

Expertise levels	self-reported game frequency	tutorial score
Beginner	lower than 10% of mean	lower than 10% of mean
Intermediate	higher than 10% of mean	lower than 10% of mean
	falls in mean - 10% and mean + 10%	falls in mean - 10% and mean + 10%
Expert	higher than 10% of mean	higher than 10% of mean
	lower than 10% of mean	higher than 10% of mean

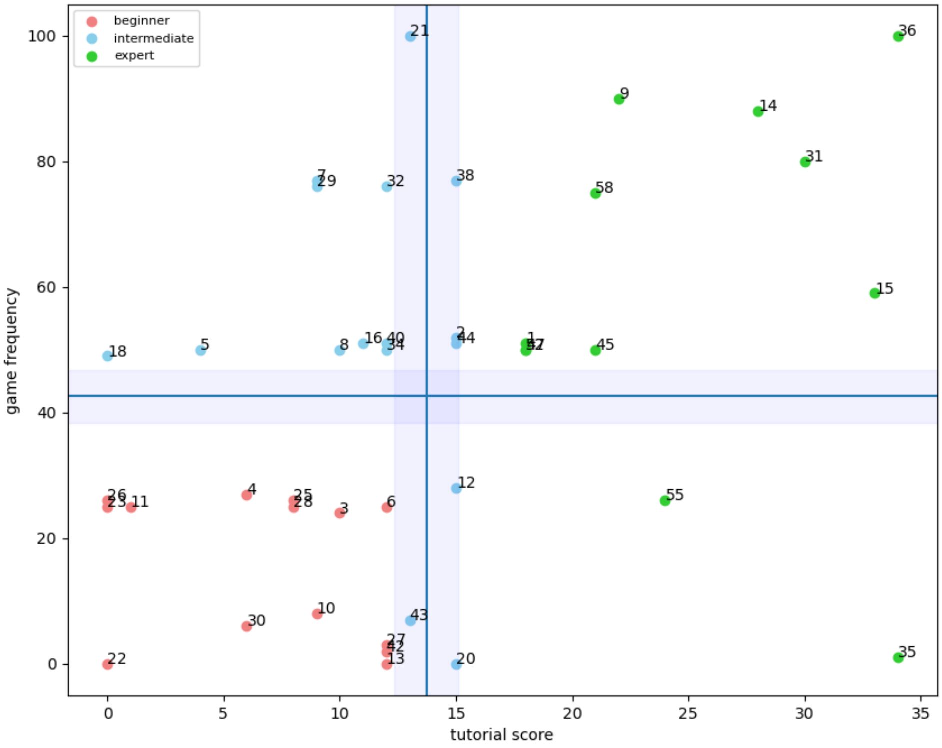


Figure 5.2: Division of different expertise levels

5.3. Effects of expertise levels and information amount

After dividing participants into three expertise groups: beginners, intermediate, and expert gamers, we can investigate the effect of expertise levels and information amount on subjective workload, task performance, trust, and explanation satisfaction. Both expertise levels and information amount are transferred into categorical variables in the data frame. For further analysis, *less info* task is coded as 0, and *more info* task is coded as 1. For three expertise levels, beginners are coded to 0, participants in the intermediate level are coded to 1, and experts are coded to 2.

Based on the conceptual model in this study, there are two independent variables, with one within-subject and one between-subject variable, a two-way mixed ANOVA or a non-parametric equivalent test is conducted according to the data distribution. In the following sections, we will first check the data distribution, and if it satisfies the requirements for ANOVA, we will conduct the ANOVA, otherwise, a non-parametric equivalent ANOVA will be performed instead.

5.3.1. Subjective workload

The distribution of workload in *less info* and *more info* tasks of three expertise levels are visualized in Figure 5.3. The distribution of subjective workload has four outlier, hence a non-parametric mixed ANOVA is performed to check the effect of information amount and expertise levels on workload. The results of ANOVA are presented in Table 5.2.

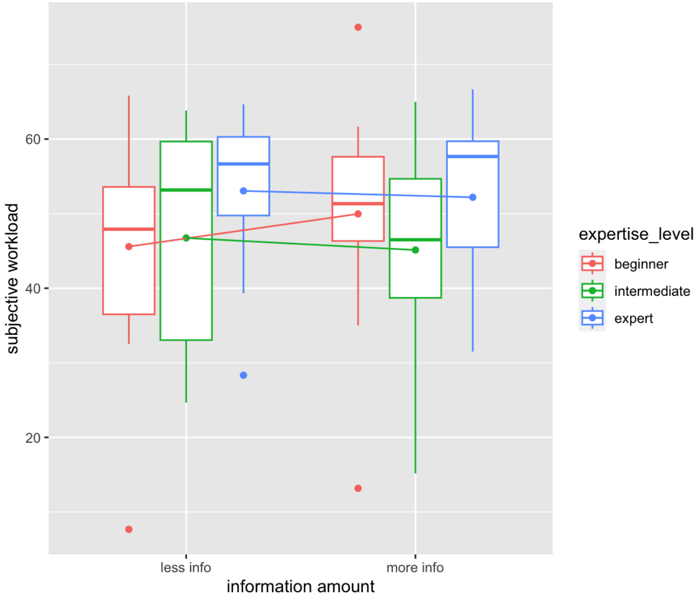


Figure 5.3: Interaction plot of the effect of information amount and expertise levels on subjective workload

Table 5.2: ANOVA result of workload

Effect	Statistic	df	p
expertise level	1.077	1.994	0.340
information amount	0.001	1.00	0.980
expertise level : information amount	1.468	1.969	0.231

5.3.2. Expertise levels and task performance

As the metric of task performance, the task score's distribution is visually presented in Figure 5.4. During the two-way mixed ANOVA assumption check, there are 8 outliers in the task score. To provide a robust result, we chose to use a non-parametric mixed ANOVA even though the data satisfied the other prerequisites. The results of ANOVA are presented in Table 5.3. Neither expertise levels nor information amount presented a significant result, and we did not find a significant effect of the interaction of these two variables. Hence, we did not perform post hoc analysis.

Table 5.3: Nonparametric mixed ANOVA result of task performance

Effect	Statistic	df	p
expertise level	2.112	1.957	0.122
information amount	0.003	1.000	0.957
expertise level : information amount	0.869	1.794	0.409

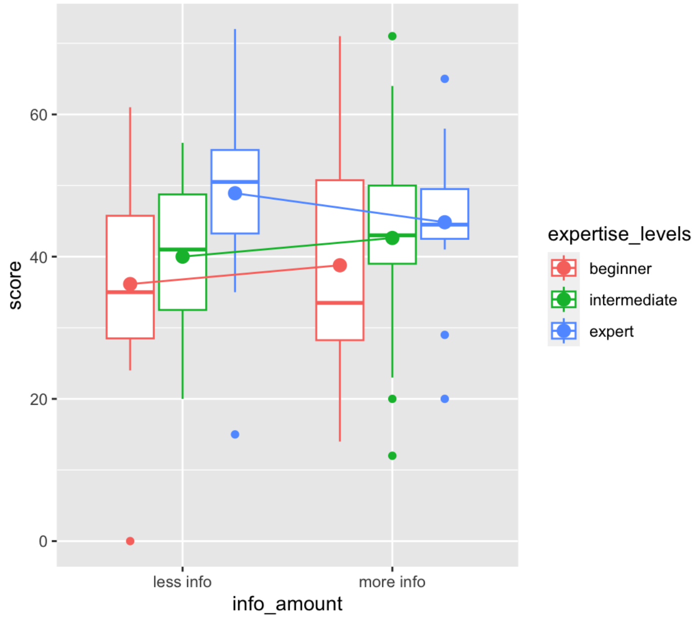


Figure 5.4: Interaction plot of the effect of information amount and expertise levels on task score

5.3.3. Expertise levels and explanation satisfaction

The distribution of explanation satisfaction is visualized in Figure 5.5. The data violated the assumption of the normality distribution in two-way mixed ANOVA and has three outliers. Hence, a non-parametric mixed ANOVA is performed instead. The results of non-parametric mixed ANOVA are presented in Table 5.4. None of the expertise level, information amount, or the interaction of these two variables has a significant effect on explanation satisfaction.

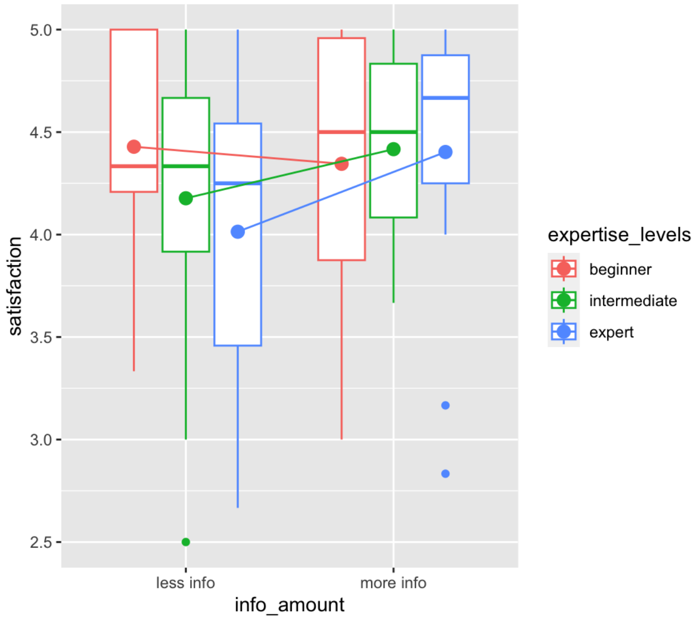


Figure 5.5: Interaction plot of the effect of information amount and expertise levels on explanation satisfaction

Table 5.4: Nonparametric mixed ANOVA result of explanation satisfaction

Effect	Statistic	df	p
expertise level	0.395	1.885	0.661
information amount	1.377	1.000	0.241
expertise level : information amount	0.965	1.782	0.373

5.3.4. Expertise levels and trust

The data distribution in trust is visualized in Figure 5.6. The data in trust violates the normality assumption of two-way mixed ANOVA, hence we use a non-parametric equivalent ANOVA to compare the effect of information amount and expertise levels on trust. The statistical results are presented in Table 5.5. Neither of the main effects of expertise level and the information amount, nor the interaction effect is significant.

Table 5.5: ANOVA result of trust

Effect	Statistic	df	p
expertise level	0.333	1.996	0.717
information amount	0.500	1.000	0.480
expertise level : information amount	2.688	1.478	0.084

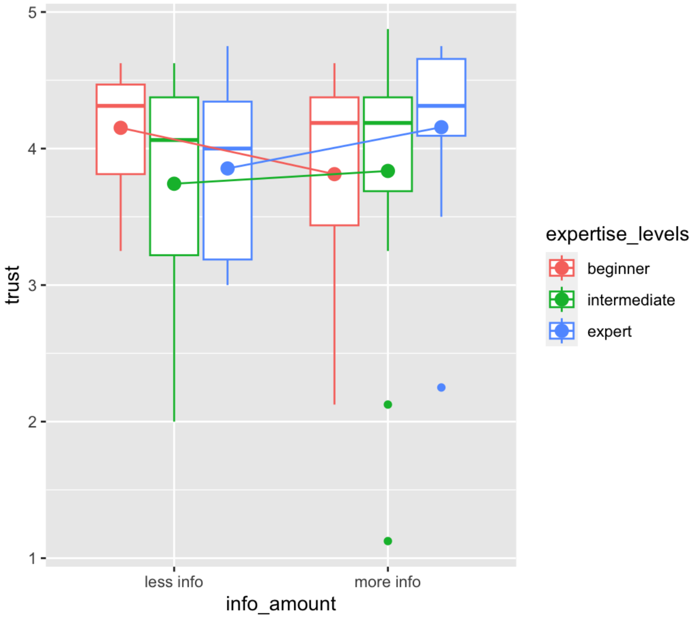


Figure 5.6: Interaction plot of the effect of information amount and expertise levels on trust

5.3.5. Activity level

After we did not find significant results in dependent variables: workload, performance, trust, and explanation satisfaction, we would like to further analyze the internal process of the experiment by analyzing the participants' total moves, idle time, and response time during the tasks. The distribution in these three variables is visualized in Figure 5.7.

Response time

The response time is calculated by how much time it takes participants to reply to the agent's messages. A non-parametric mixed ANOVA was performed to investigate the effect of information amount and expertise levels on response time. Though the results did not show a significant effect for further analysis, in Table 5.6 the p-value of the effect of expertise level on response time is marginally significant.

Table 5.6: Nonparametric mixed ANOVA result of response time

Effect	Statistic	df	p
expertise level	2.927	1.982	0.054
information amount	2.529	1.000	0.112
expertise level : information amount	0.446	1.947	0.635

Idle time

The idle time is the time when participants were either not moving or not working on an action. A non-parametric mixed ANOVA was performed to investigate the effect of information amount and expertise levels on idle time. The results did not show significant effects of information amount, expertise levels, or the interaction between these two variables.

Total moves

The total moves calculate how many steps participants walked on the map. A non-parametric mixed ANOVA was performed to investigate the effect of information amount and expertise levels on total moves. The results did not show significant effects of information amount, expertise levels, or the interaction between these two variables.

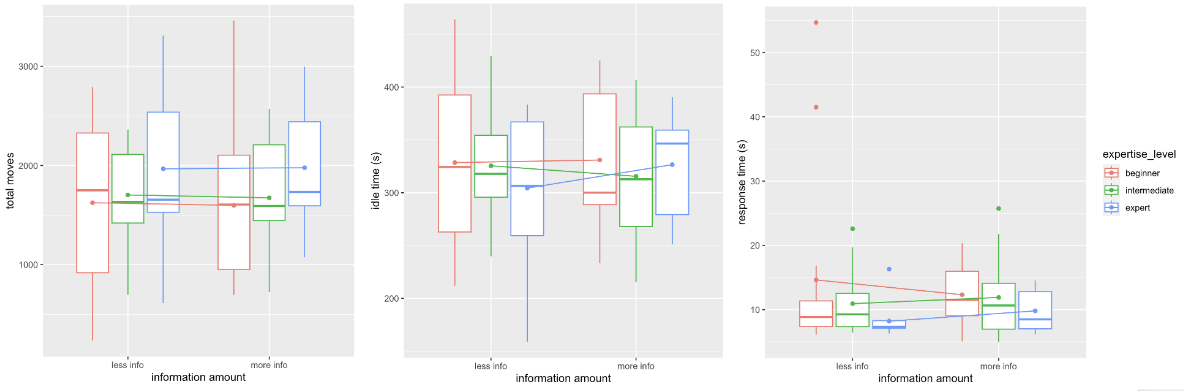


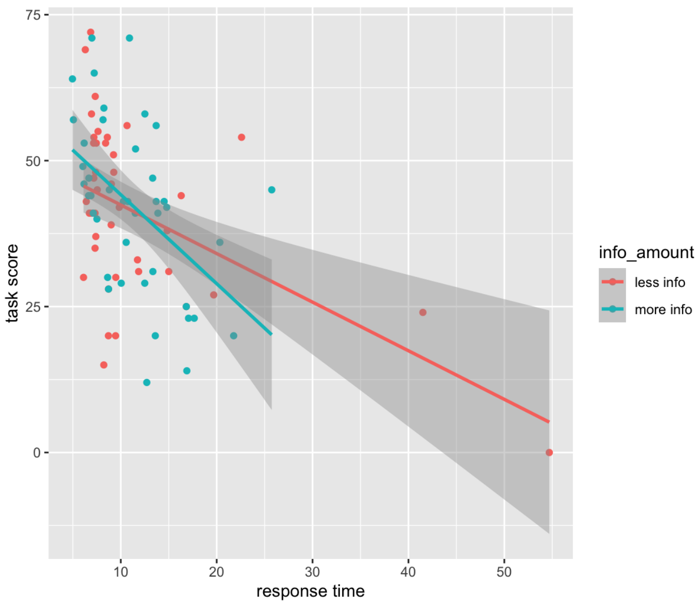
Figure 5.7: Interaction plot of the effect of information amount and expertise levels on total moves, idle time (in seconds), and response time (in seconds)

5.4. Regression analysis

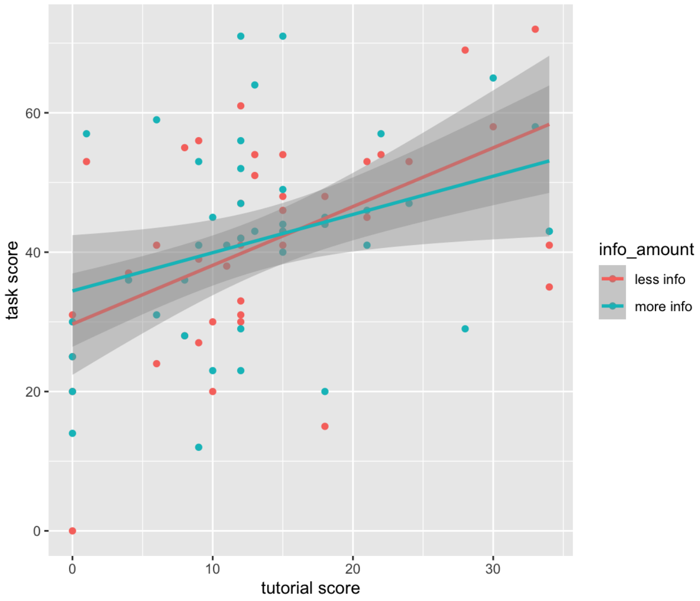
After we checked the effect of information amount and expertise levels on dependent variables, we also want to investigate whether we can predict dependent variables quantitatively by some predictors. Based on our conceptual model, we assume the possible predictor for dependent variables includes information amount, expertise levels, and workload. The information amount and expertise levels are used as categorical variables, with 2 and 3 categories respectively.

Task performance

To predict task performance, we use a multi-linear regression model that contains predictors of tutorial score, game frequency, response time, total moves, and interaction between expertise level, workload, and information amount. The multi-linear regression model satisfies all five assumptions for the model: Global statistics, skewness, kurtosis, link function, and heteroscedasticity. The model can significantly predict task performance (F-statistic=4.316, Adjusted R-squared=0.2854, p-value=9.597e-05). Besides, the response time (p-value<0.001) and tutorial score (p-value<0.001) are both significant predictors of task performance. Figure 5.8a and 5.8b are the visualizations for predicting task score by response time and tutorial score.



(a) Predicting task score by response time



(b) Predicting task score by tutorial score

Figure 5.8: Predicting task score by response time and tutorial score, the red and green lines represent the fitted regression line for *less info* and *more info* tasks, and the grey bands represent the 95% confidence interval limits.

Trust

To quantitatively predict trust, we use the same predictors as predicting task performance, which are categorical information amount and expertise levels, tutorial score, self-reported game frequency, and workload. However, the model cannot significantly predict trust (F-statistic=0.701, Adjusted R-squared=-0.029, p-value=0.689).

Explanation satisfaction

To quantitatively predict explanation satisfaction, we use the same predictors as predicting task performance, which are tutorial score, game frequency, response time, total moves, and interaction between expertise level, workload, and information amount. The multi-linear model can significantly predict trust (F-statistic=2.237, Adjusted R-squared=0.1408, p-value=0.021). Among all these predictors, workload is a significant predictor of predicting explanation satisfaction (p-value<0.001). Figure 5.9

presents the visualization to predict explanation satisfaction using the workload score.

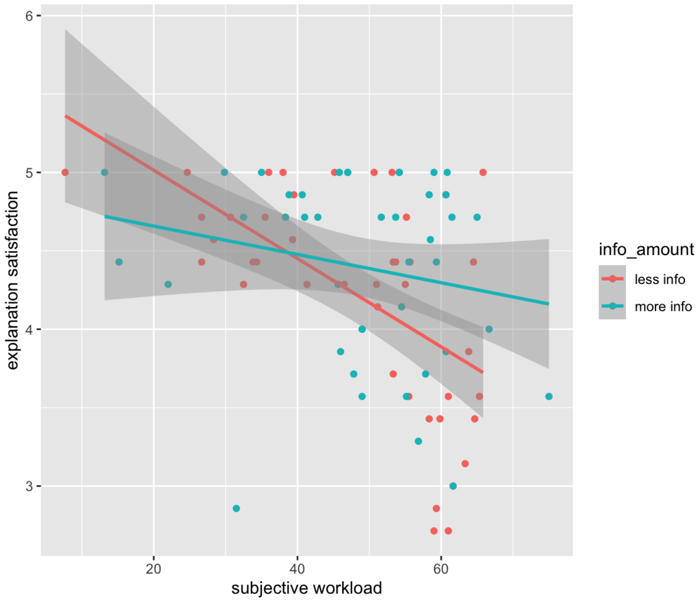


Figure 5.9: Predicting explanation satisfaction by subjective workload, the red and green lines represent fitted regression line for less info and more info tasks, the grey bands represent 95% confidence interval

5.5. Qualitative analysis

In this section, we present the answers to the open questions and the feedback from participants.

5.5.1. Preference of two agents

In Figure 5.10, we plotted a histogram of all participants' self-reported preferences for the two agents: providing explanations including less or more information. Half of the participants prefer the agent providing explanations with less information. 1/3 participants did not find any preference between the two agents. Only 1/6 of the participants prefer the agent that provides explanations with more information.

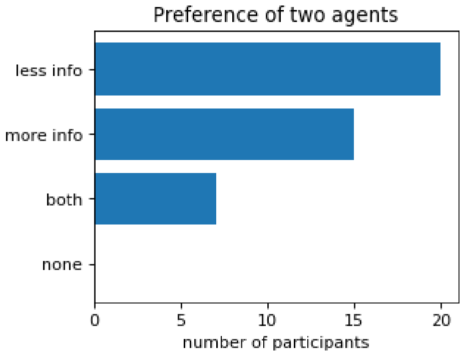
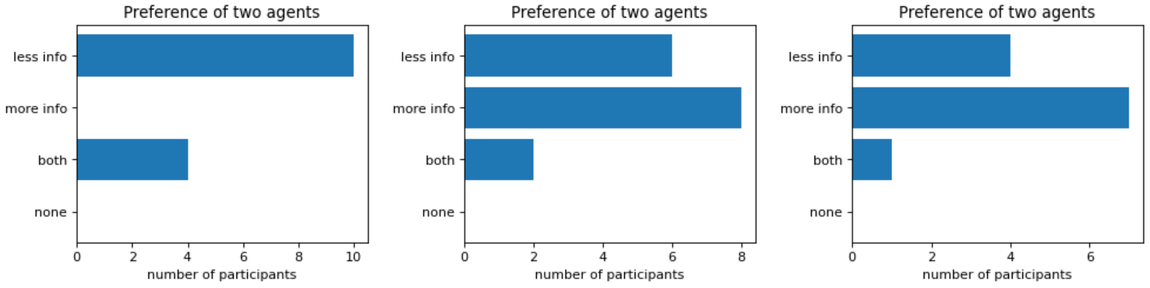


Figure 5.10: Subjective preference of all participants

To further analyze the subjective preference within three expertise levels, we separated participants' answers based on their expertise groups. In Figure 5.11a, 71.4% beginners reported a subjective preference for *less info* agent, and the other participants did not feel a preference for either of the two agents. None of the beginners prefer the *more info* agent. While in Figure 5.11b and 5.11c, most intermediate and expert gamers tend to prefer the *more info* agent, with 50% intermediate gamers and 58.3% experts reported a preference for the *more info* agent.



(a) Beginners **(b)** Intermediate **(c)** Experts

Figure 5.11: Subjective preference of three expertise levels

However, the reasons for participants' subjective preferences provided by participants are not always related to the explanations. Since participants who did not express a preference were not required to provide a reason for that, 28 participants answered the reasons for their preferences. Within these 28 participants, 42.8% of the participants provided reasons for their preferences that are not related to the explanations, such as 'I feel this robot is smarter', 'This robot moves faster', and 'I get a better score by working with this robot'.

5.5.2. Feedback to the open questions

From the open questions, we received 28 valid feedback to the agents' explanations, and 20 valid feedback to the task design in the experiment. We first extracted the keywords from the suggestions and then calculated the frequencies of these keywords. The bold texts are the most frequently mentioned keywords in the feedback. The details of the feedback are presented below.

Feedback on the explanation design

The frequency of each keyword and the feedback to the explanation design is visualized in Figure 5.12. The most mentioned suggestions are related to the length or information density of the explanations.

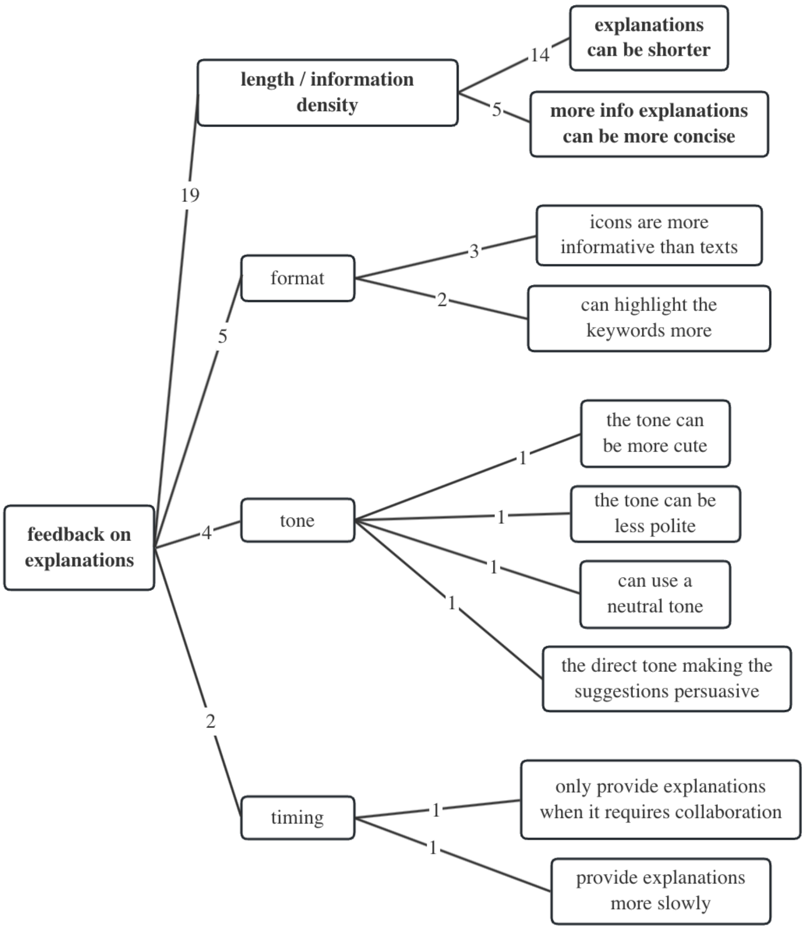

feedback on explanations
19
length / information density
14
explanations can be shorter
5
more info explanations can be more concise
5
format
3
icons are more informative than texts
2
can highlight the keywords more
4
tone
1
the tone can be more cute
1
the tone can be less polite
1
can use a neutral tone
1
the direct tone making the suggestions persuasive
2
timing
1
only provide explanations when it requires collaboration
1
provide explanations more slowly

Figure 5.12: Feedback on the explanations

Feedback on the task design

The feedback to the task design is visualized in Figure 5.13. The most frequently mentioned suggestions are related to the agent's performance. The numbers in Figure 5.13 are the frequency of each feedback being mentioned.

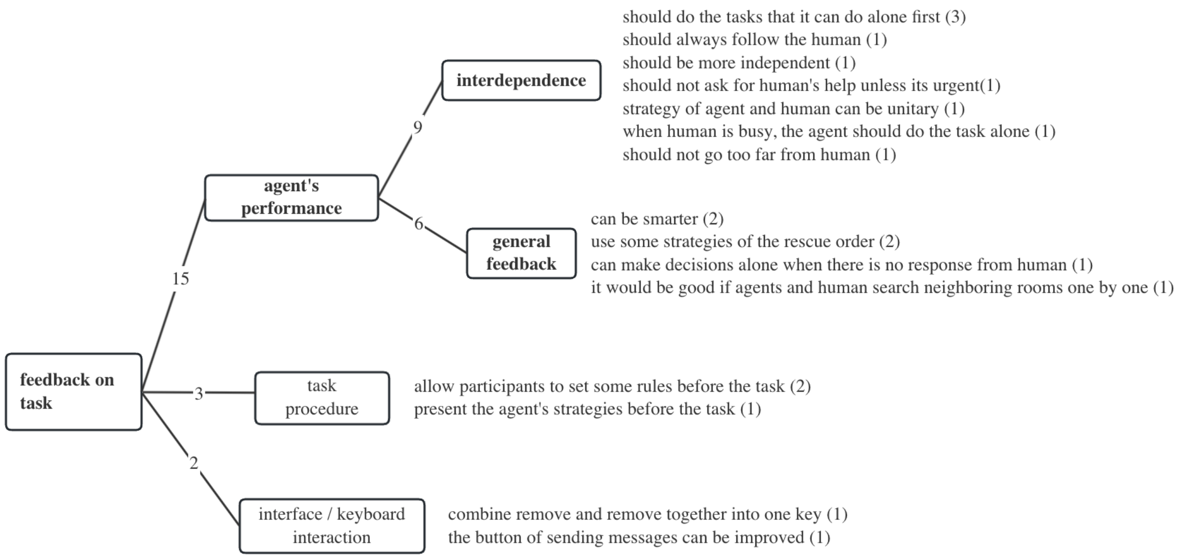

feedback on task
15
agent's performance
9
interdependence
should do the tasks that it can do alone first (3)
should always follow the human (1)
should be more independent (1)
should not ask for human's help unless its urgent(1)
strategy of agent and human can be unitary (1)
when human is busy, the agent should do the task alone (1)
should not go too far from human (1)
6
general feedback
can be smarter (2)
use some strategies of the rescue order (2)
can make decisions alone when there is no response from human (1)
it would be good if agents and human search neighboring rooms one by one (1)
3
task procedure
allow participants to set some rules before the task (2)
present the agent's strategies before the task (1)
2
interface / keyboard interaction
combine remove and remove together into one key (1)
the button of sending messages can be improved (1)

Figure 5.13: Feedback on the task

6 Discussion

In the previous section, we presented the results of the experiment. This section further discusses the results and tries to formulate an answer to the research questions.

6.1. Research question

The main research question of this study is: **what is the effect of less versus more detailed agent explanations on human-agent teamwork for people with different gaming expertise?** After we get all the results, our model can be visualized in Figure 6.1.

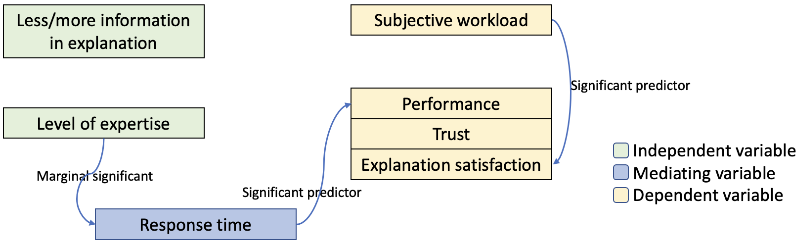


Figure 6.1: Conceptual model with results

To compare experts' and beginners' preferences, we first separated all participants into different expertise levels. In this study, the method that we divided all participants into different expertise levels is a combination of self-reported game frequency and tutorial scores. Thus, experts include two types of participants: the ones who are both frequent gamers and also performed better than average in the tutorial, and two participants with extremely good performance in the tutorial. Performing the tutorial better than the others requires faster comprehension of the rules, better knowledge of the situation, and more efficient collaboration with the agent. Though there might be other methods to divide expertise levels, given the data we collected, we believe this is a sensible way to distinguish experts and beginners from others without mixing with the actual task performance.

6.1.1. Workload

According to the results in Section 5.3.1, we did not find a significant difference in subjective workload between the two tasks and between different expertise levels. The previous hypothesis of workload in the two tasks is that the task with more info explanations is supposed to increase participants' subjective workload, but given this result, we do not have evidence that increasing the information amount in explanation also increases participants' subjective workload. This result is in line with the study by [70], [47]. However, the results of no significant difference between different expertise levels did not prove our assumption and are not in line with the results from [59], which claimed that better cognitive abilities have an effect on mental workload. We speculate that though different expertise levels have different perceptual and cognitive abilities, the amount of information in explanations is not an obvious factor to influence subjective workload in different expertise levels.

Given the task design, working on the SAR task can be divided into several sub-tasks: moving around on the map and searching for obstacles or victims; paying attention to the chat; and communicating. When participants worked on the task, every sub-task consumed some effort, so the activity level of total moves, idle time, and response time can be interpreted as an aspect of the objective workload. The ANOVA of response time presents a marginally significant effect of expertise level on response time (p-value=0.054), which is in line with the results from [47] that active video gamers have less response time than non-active video gamers in the conditions of no explanations and transparent conditions.

6.1.2. Explanation satisfaction

Comparing participants' explanation satisfaction in different expertise levels is the most straightforward way to investigate the preference of a certain expertise level for a specific type of explanation. After observing the ANOVA results of explanation satisfaction, we did not find significant differences in explanation satisfaction in either *less* or *more info* explanations or in the three expertise levels. This is different from our hypothesis, which assumed that *more info* explanations can increase satisfaction for experts, while beginners are more satisfied with *less info* explanations.

For both *less info* and *more info* tasks, we did not find a significant difference in the explanation satisfaction of all participants. To recap our explanation design, though there is a difference in the amount of information between *less info* and *more info* explanations, there is also some information overlap between the two types of explanations. No significant difference between these two types of explanations indicates a similar effect of *less info* and *more info* explanations on the tasks. However, Figure 5.5 shows a relatively high mean explanation satisfaction (more than 4.0), which indicates participants tend to find both two explanations satisfactory. Hence, we speculate that the information amount in the explanations is a subtle factor that influences explanation satisfaction, the general tone or format of explanations might have a more significant effect on satisfaction.

Besides, the study by [36] found adults are goal-driven learners and prefer goal-based explanations to belief-based explanations. In our study, we already clarified the agent's situation and suggestions in *less info* explanations as a goal, which can make participants work on the tasks. Therefore, we speculate the effect of explaining more beliefs and sub-goals of the agent may not significantly improve explanation satisfaction.

In the regression analysis, we found the task's workload can be a significant predictor for explanation satisfaction. The higher the workload participants experienced, the less satisfied they felt with the explanations. Given this result, we speculate that as part of the ability of the XAI agent, the perception and user experience of explanations can be negatively affected by participants' subjective workload. Given a relatively higher workload, participants tend to be less satisfied with the agent's explanation due to the higher pressure caused by their workload.

6.1.3. Trust

Past experience has an impact on people's trust relationships [24]. However, in this study, due to the division of expertise groups, being an expert does not equal having past experience in a similar task to the ones in our experiment. To investigate whether less or more information would influence trust during the tasks, it is necessary to measure initial trust before the study.

Based on the results, we did not find a significant difference in subjective trust between the two tasks. Hence, there is no evidence that providing more information in the explanations can help to increase subjective trust in agents, which is in line with the results from [70]. Though studies by [8] found displaying transparent information is effective for trust calibration, the effect of increasing information with the same format in explanations seems subtle in our study.

6.1.4. Performance

The only difference between the two tasks is the amount of information in the explanations provided by the agent. Since we did not find a significant difference between the performance scores of the two tasks, the amount of information in the explanations does not significantly affect the task scores. Based on the task design and explanation design, we assume providing explanations with more information can help participants know more about the agent's beliefs. Besides, with an accurate reason for the agent's suggestion, participants can react to the agent's suggestion more sensibly, which also helps to improve the team performance. Given these two advantages of *more info* explanations, and experts' better capability to process more information, we assume providing more information can help experts increase performance. However, the ANOVA results of the task performance did not present a significant effect of expertise level or information amount on task performance. We speculate that the amount of information in explanations is not the most important factor for performance. Moreover, though experts have the ability to work with more information, they may not need much information to perform a task well.

Despite the non-significant effect from ANOVA, we do find that tutorial score and response time can be significant predictors of task performance. This indicates that task performance is more related to participant's prior knowledge and their attentiveness to the task, but the information in explanations and the expertise levels do not significantly affect the task performance.

6.1.5. Game frequency

According to Figure 5.1, the self-reported game frequency does not show a correlation to the task performance during this experiment. We speculate two possibilities for this. The first speculation is that it can be because the types of video games that participants usually played are not similar to the ones in our experiment. For example, a person who frequently plays competitive video games might not be good at cooperating with others in a task. The second speculation is that there is a bias between subjective game frequency and objective game performance. As we discussed in Section 2.2.2, frequently playing games does not necessarily make one an expert, moreover, there is a bias between humans' subjective opinions and objective performance.

6.1.6. Qualitative analysis

In the open questions of preference between the two agents, the most mentioned keywords are short and concise. However, there is a difference between short and concise: a short message might not contain sufficient information, while a concise message keeps the length as short as possible with sufficient information. Though short does not necessarily mean concise, it is noted that length is the most frequently mentioned aspect of how participants evaluate their preference for an explanation. Longer explanations will not make participants satisfied and willing to read, which also conforms with the studies [37], [52] that argue a good explanation is supposed to be short. Hence it is important to shorten the explanations while keeping the same information.

One thing to notice is that Figure 5.11a, 5.11b and 5.11c, present a difference in the number of participants in three expertise levels reported their subjective preference for *less info* and *more info* agents, this is contradictory to the results in Section 5.3.3 that there is no effect of information amount and expertise levels on explanation satisfaction. One possible reason is that there is a difference in measuring explanation satisfaction by two separate questionnaires and explicitly asking for participants' preferences. An explicit question like "Which agent do you prefer" implies that there is a difference between the two agents, and participants need to choose one of them. Hence, though we named the agents "SaR bot1" and "Sar bot 2", and provided options like "both agents are fine", "I prefer none of the agents", this question may still lead participants to recall and think about the difference between the two agents, which could lead to a biased result.

6.2. Limitations

There are several limitations in this study, and it can be divided based on the stages of the study. Firstly, in the pilot study, the self-reported game experts did not reach a relatively high score in the tasks. The ideal situation would be that self-reported experts can provide some suggestions on the explanation design that is more useful for experts. Since the self-reported experts did not perform the task with an expected score, the suggestions provided by these participants might not be useful in adjusting the explanations to the experts. Hence, the number of recruited experts in the pilot study can be increased to collect more suggestions to adjust the more info explanations.

In our experiment, the two conditions that we compared are *less info* and *more info*, so given the non-significant results of ANOVA, we can only conclude that we did not find evidence for the influence of the amount of information in explanations. Another limitation in the experiment design is that we did not add uncertainties in the agent, while some studies, like [11], investigated the effect of transparency level on trust always involving uncertainty in the agent's explanations. In our research, to make the *more info* explanations more helpful, we could also add some faults or uncertainties in our agent's behavior.

In the context of this study, participants were required to perform tasks in a limited time. From the participant's perspective in this study, accomplishing the task consists of certain types of effort, such as reading explanations, memorizing the map, deciding on the next action, etc. Completing the task with a relatively high score requires all these types of effort, but different participants may have different methods to allocate their effort and time. Some participants may spend more time reading the explanations carefully, while some participants may spend more effort memorizing the map. All these efforts compose their workload during the task. Though interpreting workload into different dimensions is sensible, measuring workload by different sub-tasks can help to interpret workload as well.

There are also some limitations in the search and rescue game we designed. According to the feedback from participants, the keyboard actions can be more convenient and easy to remember. Although we presented the instructions for the keyboard actions at the top of the interface, participants responded that it was hard for them to look at the keyboard instructions every time. From the perspective of task setting, though we have emphasized this search and rescue task needs participants to work with the agent as a team, some participants still tend to work competitively and independently with the agent. In this study, we did not measure the subjective or objective collaboration level of participants. Besides, the task in this study is designed with relatively high independence for human participants, which means humans can also perform the task without interacting with the agent.

From the perspective of explanation design, explanations can be limited in this study since they were considered based on the search and rescue task scenario. However, explanation length is the independent variable of this study, and explanations can be designed based on the hypotheses themselves. The search and rescue task can be designed based on the explanations. From the explanation modality perspective, although we tried to highlight the key information through images, the explanations in this study are provided only by text. Some participants mentioned it could be easier for them to work on the task if the explanations could be provided by audio.

6.3. Future study

For future studies, the first thing that can be improved is the measurement of self-reported game frequency. In the results, we did not find a correlation between self-reported game frequency and task performance. This can be caused by the measurement of self-reported game frequency, or there is no correlation between these two variables. In this study, we used the question of *"How often do you play video games"* to measure subjective self-reported game frequency, without mentioning specific types of video games or how they would rate their own gaming expertise levels. For future studies, we suggest the question of game frequency or game expertise can be specified by emphasizing certain types of video games. People who are good at playing a specific type of video game can still be good at a similar game but are not necessarily good at a different type of game. Hence, it can be important to mention the type of video games to further investigate the correlation between self-reported and objective game expertise.

In the task design, we chose to display the explanations and the task at the same time. If there is a need to force participants to read the explanations, future studies can implement a pop-up window to display the explanations as well. However, from the teaming perspective, forcing participants to do something might reduce the user experience. Hence, designing a task that both achieves good teamwork and makes participants pay sufficient attention to the explanations can require a delicate balance between task design and explanation design.

From the ANOVA results of trust, the p-value for the interaction effect of expertise level and information amount on trust is marginally significant (p-value=0.084). Though we cannot further decompose the interaction effect due to a non-significant (p-value>0.05) result, we speculate there might be an interaction effect but given the data we collected, the results are not significant.

In the previous discussion, we speculated that one of the reasons that we did not find a significant effect of expertise levels on those dependent variables is that experts may not need too much information to achieve a good performance. Hence, future studies can investigate whether this speculation is fair. Moreover, based on the insights from the experiment and feedback, it might be interesting to investigate how different sentence patterns and tones influence explanation satisfaction. Though there are already results stating shorter explanations are preferred [52], it might be interesting to investigate whether longer explanations are not preferred because of first impressions of the longer texts, a much more cognitive workload, or is caused by the time pressure under a time-limited task.

7 Conclusion

In this study, we designed and conducted a simulated 2D game-like search and rescue task, and investigated the effect of less or more detailed explanations and three expertise levels on subjective workload, task performance, trust, and explanation satisfaction. Based on the division method using self-reported game frequency and the mock task score in the tutorial, we divided 42 participants into 3 expertise levels: 14 beginners, 16 intermediate gamers, and 12 expert gamers. Given this division, we analyzed the effect of less and more information explanation and expertise levels on subjective workload, team performance, trust, and explanation satisfaction. The mixed ANOVA results did not show any significant difference between or within the groups, which is not consistent with our hypothesis that providing more information in explanations helps experts improve performance, trust, and explanation satisfaction.

Our hypothesis is based on the theory that experts with higher cognitive ability and better attention are better at processing more information while doing the task at the same time. Providing more information may help to increase humans' belief in the agent, hence improving explanation satisfaction. Given the result of no significant difference, we speculate that although expert gamers have better abilities in attentiveness and cognition, generally experts may not require too much information in explanations during a simulated human-agent teamwork task. Especially in this study, where actions require soft interdependence more than hard interdependence, participants have plenty of extents to work alone, and *less info* explanations already provide the most necessary information for teamwork, such as location, and suggestions, which reduces the need for experts to request more information. Besides, from the perspective of explanation satisfaction, experts may have a similar preference to the majority of participants even though they performed better in the tasks.

However, in regression analysis, we do find that participant's response time and tutorial score can be significant predictors of task performance. Besides, workload can be a significant predictor for explanation satisfaction, with a negative correlation.

To answer our research question, we cannot conclude whether increasing the amount of information in explanations helps to increase experts' experience in HAT. In a time-limited HAT task that requires operations from humans, we did not find evidence that providing more information in explanations affects trust, explanation satisfaction, or performance for people with different gaming expertise levels. However, workload is found to have a negative effect on explanation satisfaction. Hence, for future studies, it may be worth investigating whether expert gamers really need more information in the explanation of the task, or it is the opposite that expert gamers can still perform the task without too much information.

References

[1] Irene Abi-Zeid and John R Frost. "SARPlan: A decision support system for Canadian Search and Rescue Operations". In: *European Journal of Operational Research* 162.3 (2005), pp. 630–653.

[2] Ulf Ahlstrom and Ferne J Friedman-Berg. "Using eye movement activity as a correlate of cognitive workload". In: *International journal of industrial ergonomics* 36.7 (2006), pp. 623–636.

[3] Sule Anjomshoae et al. "Explainable agents and robots: Results from a systematic literature review". In: *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems. 2019, pp. 1078–1088.

[4] Stavros Antifakos et al. "Towards improving trust in context-aware systems by displaying system confidence". In: *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. 2005, pp. 9–14.

[5] Kira Bailey, Robert West, and Craig A Anderson. "A negative association between video game experience and proactive cognitive control". In: *Psychophysiology* 47.1 (2010), pp. 34–42. DOI: `https://doi.org/10.1111/j.1469-8986.2009.00925.x`.

[6] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. "The role of domain expertise in trusting and following explainable AI decision support systems". In: *Journal of Decision Systems* (2021), pp. 1–29. DOI: `https://doi.org/10.1080/12460125.2021.1958505`.

[7] Walter R Boot, Daniel P Blakely, and Daniel J Simons. "Do action video games improve perception and cognition?" In: *Frontiers in psychology* 2 (2011), p. 226. DOI: `https://doi.org/10.3389/fpsyg.2011.00226`.

[8] Michael W Boyce et al. "Effects of agent transparency on operator trust". In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. 2015, pp. 179–180.

[9] Jeffrey M Bradshaw, Paul Feltovich, and Matthew Johnson. "Human-agent interaction". In: *Handbook of human-machine interaction* (2017), pp. 283–302.

[10] Brad Cain. "A review of the mental workload literature". In: *DTIC Document* (2007).

[11] Jessie YC Chen et al. "Situation awareness-based agent transparency and human-autonomy teaming effectiveness". In: *Theoretical issues in ergonomics science* 19.3 (2018), pp. 259–282.

[12] Ana Cristina Costa, Robert A Roe, and Tharsi Taillieu. "Trust within teams: The relation with performance effectiveness". In: *European journal of work and organizational psychology* 10.3 (2001), pp. 225–244.

[13] Xinyue Dai et al. "Counterfactual explanations for prediction and diagnosis in XAI". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022, pp. 215–226.

[14] Sandra Devin and Rachid Alami. "An implemented theory of mind to improve human-robot shared plans execution". In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2016, pp. 319–326. DOI: `10.1109/HRI.2016.7451768`.

[15] Mica R Endsley and W Jones. "Situation awareness". In: *The Oxford handbook of cognitive engineering* 1 (2013), pp. 88–108.

[16] Courtney Ford and Mark T Keane. "Explaining Classifications to Non Experts: An XAI User Study of Post Hoc Explanations for a Classifier When People Lack Expertise". In: *arXiv preprint arXiv:2212.09342* (2022). DOI: `https://doi.org/10.48550/arXiv.2212.09342`.

[17] John R Frost and Lawrence D Stone. "Review of search theory: Advances and applications to search and rescue decision support". In: (2001).

[18] Ze Gong and Yu Zhang. "Behavior explanation as intention signaling in human-robot teaming". In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2018, pp. 1005–1011.

[19] Michael A Goodrich et al. "Supporting wilderness search and rescue using a camera-equipped mini UAV". In: *Journal of Field Robotics* 25.1-2 (2008), pp. 89–110.

[20] C Shawn Green and Daphne Bavelier. "Learning, attentional control, and action video games". In: *Current biology* 22.6 (2012), R197–R206. DOI: `https://doi.org/10.1016/j.cub.2012.02.012`.

[21] Stephen Grossberg. "A path toward explainable AI and autonomous adaptive intelligence: deep learning, adaptive resonance, and models of perception, emotion, and action". In: *Frontiers in Neurorobotics* 14 (2020), p. 36. DOI: `https://doi.org/10.3389/fnbot.2020.00036`.

[22] Riccardo Guidotti. "Counterfactual explanations and how to find them: literature review and benchmarking". In: *Data Mining and Knowledge Discovery* (2022), pp. 1–55.

[23] David Gunning et al. "XAI—Explainable artificial intelligence". In: *Science robotics* 4.37 (2019), eaay7120.

[24] Feyza Merve Hafizoğlu and Sandip Sen. "Understanding the influences of past experience on trust in human-agent teamwork". In: *ACM Transactions on Internet Technology (TOIT)* 19.4 (2019), pp. 1–22.

[25] Nader Hanna and Deborah Richards. "The impact of communication on a human-agent shared mental model and team performance". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 2014, pp. 1485–1486.

[26] Satoshi Hara et al. "Feature Attribution As Feature Selection". In: (2018).

[27] Sandra G Hart. "NASA-task load index (NASA-TLX); 20 years later". In: *Proceedings of the human factors and ergonomics society annual meeting*. Vol. 50. 9. Sage publications Sage CA: Los Angeles, CA. 2006, pp. 904–908.

[28] Robert R Hoffman et al. "Metrics for explainable AI: Challenges and prospects". In: *arXiv preprint arXiv:1812.04608* (2018).

[29] Ashley M Hughes et al. "Cardiac measures of cognitive workload: a meta-analysis". In: *Human factors* 61.3 (2019), pp. 393–414.

[30] Vidhi Jain et al. "Predicting human strategies in simulated search and rescue task". In: *arXiv preprint arXiv:2011.07656* (2020).

[31] Matthew Johnson and Alonso Vera. "No AI is an island: the case for teaming intelligence". In: *AI magazine* 40.1 (2019), pp. 16–28. DOI: `https://doi.org/10.1609/aimag.v40i1.2842`.

[32] Matthew Johnson et al. "Autonomy and interdependence in human-agent-robot teams". In: *IEEE Intelligent Systems* 27.2 (2012), pp. 43–51.

[33] Matthew Johnson et al. "Autonomy and interdependence in human-agent-robot teams". In: *IEEE Intelligent Systems* 27.2 (2012), pp. 43–51. DOI: `10.1109/MIS.2012.1`.

[34] Matthew Johnson et al. "Coactive design: Designing support for interdependence in joint activity". In: *Journal of Human-Robot Interaction* 3.1 (2014), pp. 43–69. DOI: `https://doi.org/10.5898/JHRI.3.1.Johnson`.

[35] Matthew Johnson et al. "The fundamental principle of coactive design: Interdependence must shape autonomy". In: *Coordination, Organizations, Institutions, and Norms in Agent Systems VI: COIN 2010 International Workshops, COIN@ AAMAS 2010, Toronto, Canada, May 2010, COIN@ MALLOW 2010, Lyon, France, August 2010, Revised Selected Papers*. Springer. 2011, pp. 172–191.

[36] Frank Kaptein et al. "Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults". In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2017, pp. 676–682. DOI: `10.1109/ROMAN.2017.8172376`.

[37] Frank C Keil. "Explanation and understanding". In: *Annu. Rev. Psychol.* 57 (2006), pp. 227–254.

[38] Avi Knoll et al. "Measuring cognitive workload with low-cost electroencephalograph". In: *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part IV 13*. Springer. 2011, pp. 568–571.

[39] Ramaravind Kommiya Mothilal et al. "Towards unifying feature attribution and counterfactual explanations: Different means to the same end". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 652–663.

[40] Retno Larasati, Anna De Liddo, and Enrico Motta. "The effect of explanation styles on user's trust". In: (2020).

[41] Andrew J Latham, Lucy LM Patston, and Lynette J Tippett. *Just how expert are "expert" video-game players? Assessing the experience and expertise of video-game players across "action" video-game genres*. 2013. DOI: `https://doi.org/10.3389/fpsyg.2013.00941`.

[42] Brian Y Lim and Anind K Dey. "Design of an intelligible mobile context-aware application". In: *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*. 2011, pp. 157–166.

[43] Jordan Litman. "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance". In: (2023).

[44] Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press, 2006.

[45] John E Mathieu et al. "The influence of shared mental models on team process and performance." In: *Journal of applied psychology* 85.2 (2000), p. 273.

[46] Nathan McNeese et al. "Understanding the role of trust in human-autonomy teaming". In: (2019).

[47] Joseph E Mercado et al. "Intelligent agent transparency in human–agent teaming for Multi-UxV management". In: *Human factors* 58.3 (2016), pp. 401–415.

[48] Tim Miller. "Contrastive explanation: A structural-model approach". In: *The Knowledge Engineering Review* 36 (2021), e14.

[49] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38. DOI: `https://doi.org/10.1016/j.artint.2018.07.007`.

[50] Grégoire Milliez et al. "Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring". In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2016, pp. 43–50. DOI: `10.1109/HRI.2016.7451732`.

[51] Yazan Mualla et al. "The quest of parsimonious XAI: A human-agent architecture for explanation formulation". In: *Artificial Intelligence* 302 (2022), p. 103573. DOI: `https://doi.org/10.1016/j.artint.2021.103573`.

[52] Mark A Neerincx et al. "Using perceptual and cognitive explanations for enhanced human-agent team performance". In: *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer. 2018, pp. 204–214. DOI: `https://doi.org/10.1007/978-3-319-91122-9_18`.

[53] Adam C Oei and Michael D Patterson. "Enhancing cognition with video games: a multiple game training study". In: *PloS one* 8.3 (2013), e58546.

[54] Mayada Oudah et al. "How AI wins friends and influences people in repeated games with cheap talk". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018. DOI: `https://doi.org/10.1609/aaai.v32i1.11486`.

[55] Rohan Paleja et al. "The utility of explainable ai in ad hoc human-machine teaming". In: *Advances in neural information processing systems* 34 (2021), pp. 610–623.

[56] Vytautas Petrauskas et al. "XAI-based medical decision support system model". In: *Int. J. Sci. Res. Publ* 10.12 (2020), pp. 598–607.

[57] Lara Quijano-Sanchez et al. "Make it personal: a social explanation system applied to group recommendations". In: *Expert Systems with Applications* 76 (2017), pp. 36–48. DOI: `https://doi.org/10.1016/j.eswa.2017.01.045`.

[58] Susana Rubio et al. "Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods". In: *Applied psychology* 53.1 (2004), pp. 61–86.

[59] Susana Rubio-Valdehita et al. "Effects of task load and cognitive abilities on performance and subjective mental workload in a tracking task". In: *Anales de psicologıa* 28.3 (2012), pp. 986–995.

[60] Wojciech Samek and Klaus-Robert Müller. "Towards explainable artificial intelligence". In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), pp. 5–22.

[61] Julie Shah, Been Kim, and Stefanos Nikolaidis. "Human-inspired techniques for human-machine team planning". In: *2012 AAAI Fall Symposium Series*. 2012.

[62] Donghee Shin. "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI". In: *International Journal of Human-Computer Studies* 146 (2021), p. 102551.

[63] Keng Siau and Weiyu Wang. "Building trust in artificial intelligence, machine learning, and robotics". In: *Cutter business technology journal* 31.2 (2018), pp. 47–53.

[64] Maarten Sierhuis et al. "Human-agent teamwork and adjustable autonomy in practice". In: (2003).

[65] Ian Spence and Jing Feng. "Video games and spatial cognition". In: *Review of general psychology* 14.2 (2010), pp. 92–104.

[66] Katia Sycara and Michael Lewis. "Integrating intelligent agents into human teams." In: (2004). DOI: https://doi.org/10.1037/10690-010.

[67] Katia Sycara and Gita Sukthankar. "Literature review of teamwork models". In: *Robotics Institute, Carnegie Mellon University* 31 (2006), p. 31.

[68] Pilar Toril, José M Reales, and Soledad Ballesteros. "Video game training enhances cognition of older adults: a meta-analytic study." In: *Psychology and aging* 29.3 (2014), p. 706.

[69] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. "A two-dimensional explanation framework to classify ai as incomprehensible, interpretable, or understandable". In: *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer. 2021, pp. 119–138.

[70] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. "The influence of interdependence and a transparent or explainable communication style on human-robot teamwork". In: *Frontiers in Robotics and AI* 9 (2022), p. 993997.

[71] Jasper van der Waa et al. "Evaluating XAI: A comparison of rule-based and example-based explanations". In: *Artificial Intelligence* 291 (2021), p. 103404. DOI: https://doi.org/10.1016/j.artint.2020.103404.

[72] Sonia Waharte and Niki Trigoni. "Supporting search and rescue operations with UAVs". In: *2010 international conference on emerging security technologies*. IEEE. 2010, pp. 142–147.

[73] Arlette van Wissen et al. "Human–agent teamwork in dynamic environments". In: *Computers in Human Behavior* 28.1 (2012), pp. 23–33.

[74] Feiyu Xu et al. "Explainable AI: A brief survey on history, research areas, approaches and challenges". In: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer. 2019, pp. 563–574.

[75] Guang Yang, Qinghao Ye, and Jun Xia. "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond". In: *Information Fusion* 77 (2022), pp. 29–52.

[76] John Yen et al. "Agents with shared mental models for enhancing team decision makings". In: *Decision Support Systems* 41.3 (2006), pp. 634–653.

[77] Mark S Young et al. "State of science: mental workload in ergonomics". In: *Ergonomics* 58.1 (2015), pp. 1–17.

[78] Quan-shi Zhang and Song-Chun Zhu. "Visual interpretability for deep learning: a survey". In: *Frontiers of Information Technology & Electronic Engineering* 19.1 (2018), pp. 27–39.

A
Tutorial contents

Hello, my name is SaR bot. In this tutorial, I will introduce some rules for this search and rescue task. In this world, you can move with the keys W-A-S-D or the arrow keys. You can try to walk around with these keys, and once you are ready for the next tip, you can click the Ready button.

Due to our capability, we can only see each other within 3 blocks. You can try to move around, and get to know this capability. If you have understood this, and you are ready for the next tip, you can click the Ready button.

This world is separated by different rooms, which are surrounded by walls. Normal rooms are surrounded by walls, and the safe zone is surrounded by safe zone walls. During this task, we will refer to different rooms by their room number. You can go to the sign of A1. If you arrive at A1, you can click the Ready button.

Now we are standing in front of the entry of room A1. There is a rock blocks the entry of this room. You can only visualize the obstacles within 1 block. There are 3 types of obstacles in this world: small rocks, trees, and large rocks. If you are ready for the next tip, you can click the Ready button.

Now we can try to remove this rock. Both of us are able to remove rock alone. We can only remove obstacles within 1 block away. You can try to remove the rock alone by pressing R. If the removal is successful, you will see the rock disappear after a few seconds. If you have removed the rock, you can click the Ready button.

After removing this rock, we can enter room A1. You can try to search for victims in this room. There are three types of victims in this world: healthy victims, injured victims, and critical victims. By successfully rescuing, 1, 3, 6 points will be assigned to each type of victim. If you have found that healthy victim, you can click the Ready button.

Successfully rescuing a victim means carrying that victim, and then dropping that victim to the safe zone. We are able to carry healthy victims and injured victims alone. But for critical victims, we can only carry them together. If you understand everything, you can click the Ready button.

You can press G to grab victims alone. It takes some time to grab. Once finished, healthy victims will disappear. Then you can carry this victim to the safe zone. Which is surrounded by walls. If you understand everything, you can click the Ready button.

Now we can try to carry this healthy victim to the safe zone. In this world, victims are considered to be successfully rescued once they are carried to the safe zone. Once you arrive safe zone, you can click the Ready button.

You can drop the victim by pressing Q. When dropping, the victim will be dropped at the same location as yours. After you drop this victim, you can come to the entry of A2. We will try to remove an obstacle together. If you have arrived at A2, you can click the Ready button.

There is a tree that blocks the entry of A2. I am not able to remove the tree. So I suggest you help me remove it together. If you agree to work together, you can click Ok button to indicate your decision.

Once you are within one block of the obstacles, you can press P to remove an obstacle with me. If you see the tree has been removed, you can click Ready.

There is a critical victim in A2. We can only carry critical victims together. I am at the same location as the critical victim, If you come to my location, and press key C, you can see our icons changed. After that, you can click Ready.

Once we are carrying victims together, you will take charge of planning the path to the safe zone. Now you can move to the safe zone, and drop this victim at the safe zone. If you successfully dropped that critical victim, you can click the Ready button.

There is a large rock blocking the entry of A3. You cannot remove large rocks by yourself. So if you are within 1 block with any obstacle, you can request my help by clicking the Help button. I will go to your location, and if I sense any type of obstacle or victims to work together, I will wait for your keyboard instructions to work together. Now you can go to the entry of A3 and try to request Help.

Now I have moved to your location. And I sensed there was a large rock that we could work together. If you want to remove this obstacle together, you can press P. If the large rock has been removed, you can click the Ready button.

Now I have introduced all possible actions. We can go to the simulation zone to practice with the mock task before we actually start. If you are Ready for the mocking task, you can click the Ready button and remove the rock between the blue walls.

We have 3 minutes to work on this simulation task. Once the remaining time is less than 3 minutes, I will send a countdown message every minute in the chat. If you understand everything, you can click the Ready button, and start working on the task.

B
Informed Consent Form

Dear participant,

You are invited to an experiment titled "Adapting explanation to game experience in the context of human and agent teamwork". This study is being done by master's student Jing Zhou under Interactive Intelligence group, TU Delft.

The purpose of this experiment is to investigate human-agent interaction in a teamwork form, and will take you approximately 30 minutes to complete. The data will be used for the master thesis about game experience and explainable AI in human-agent teamwork. During this experiment, we will be asking you to complete a game in a 2D grid world based on a search and rescue task. First, we will ask you to fill in some questions regarding basic information, like age, gender, education level, etc. Then, we will start working on the simulated search and rescue task. The task consists of three parts: tutorial, first sub-task, and second sub-task. In this task, you are required to control the human agent, and collaborate with the search and rescue agent. As a team, your goal is to rescue as many victims as possible. After the task is completed, you will answer some questions about how you feel about the agent communication during the task. The data collected in this task is going to be used for further analysis of the thesis project of adapting explanation style to game experience.

To our best ability, we will keep your data confidential and anonymous. We will only share the data with 4TU repository in an anonymous format. During the experiment, we only collect age range, gender and education level, so your data is considered unidentifiable. The data management plan has been consulted with the data steward of TU Delft.

Your participation in this study is entirely voluntary, and you can withdraw it at any time. If you have any questions about or considerations, you are free to ask now, or send an email to j.zhou-15@student.tudelft.nl after this experiment.

1. I have read and understood the study information or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.

3. I understand that taking part in the study involves participation of a game-like search and rescue task in a 2D world, and answering questions from the questionnaire.

4. I understand that taking part in the study involves collecting the gender, age range, game frequency, and educational level. I understand that these data will be anonymized and will be stored securely to mitigate the risk of reidentifying, and the risk of data breach.

5. I understand that anonymous data (gender, age range, game frequency, educational level, task performance, and answers to the questionnaire) will be shared with 4TU repository only to be used for future research.

6. I understand that personal information collected about me will not be shared beyond the study team.

7. I understand that the identifiable personal data I provide will be destroyed after this thesis project.

8. I understand that after the research study the de-identified information I provide will be used for analysis, and the results will be published in a master thesis.

Name of participant:

Signature and Date

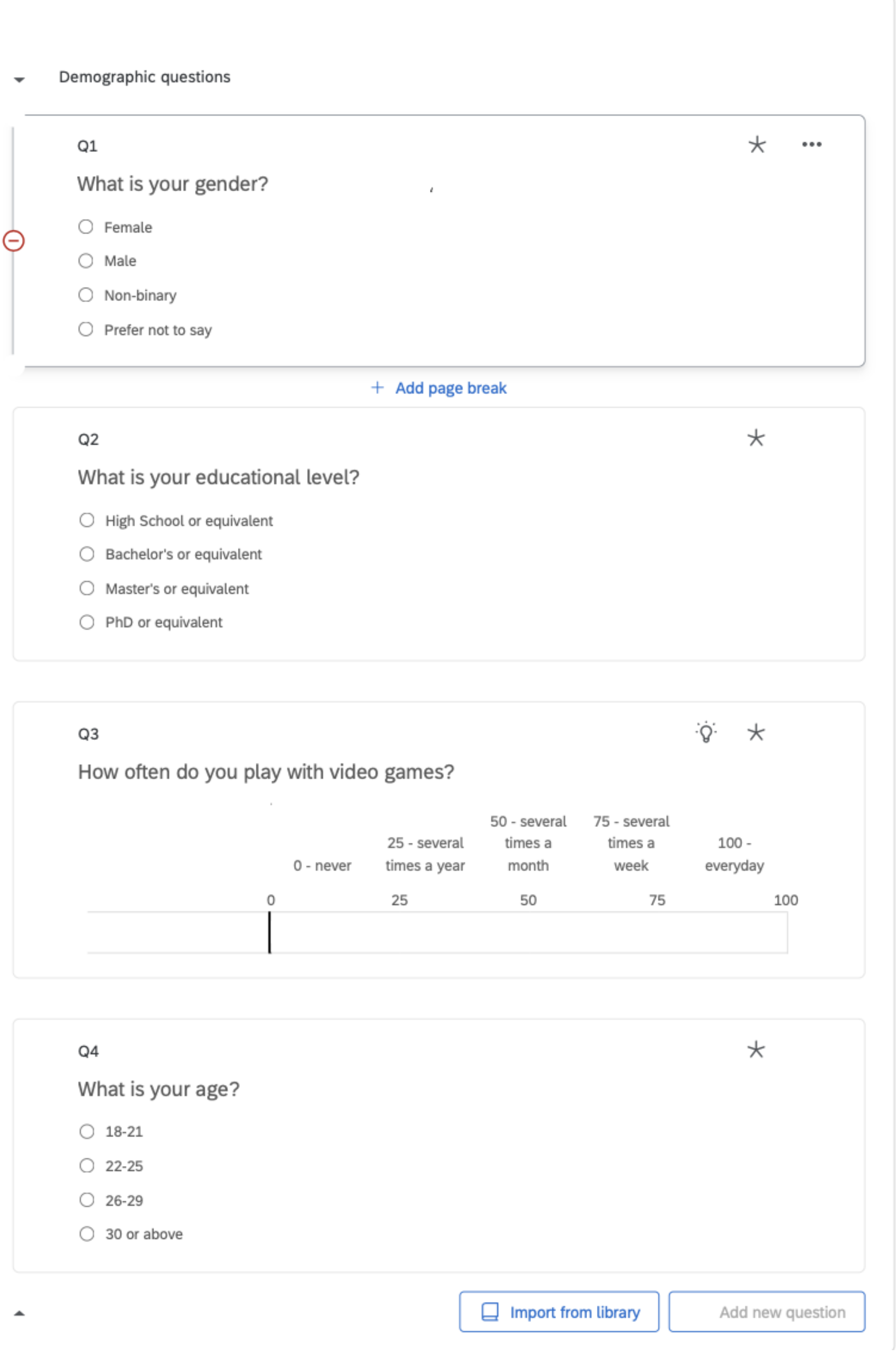
I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

Jing Zhou

Signature and Date

C

Questionnaire



Demographic questions

Q1

What is your gender?

○ Female
○ Male
○ Non-binary
○ Prefer not to say

Q2

What is your educational level?

○ High School or equivalent
○ Bachelor's or equivalent
○ Master's or equivalent
○ PhD or equivalent

Q3

How often do you play with video games?

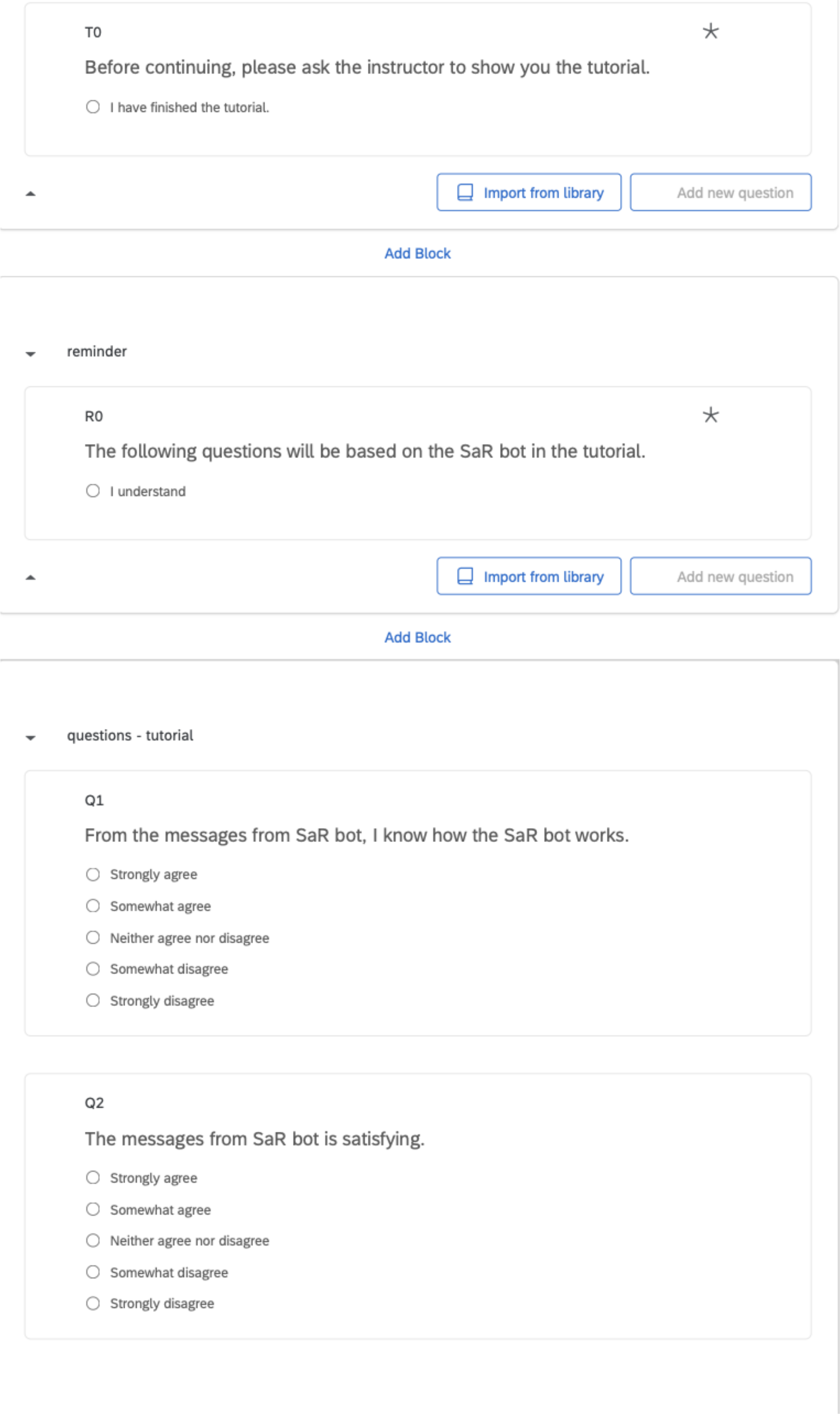
0 - never	25 - several times a year	50 - several times a month	75 - several times a week	100 - everyday

0 25 50 75 100

Q4

What is your age?

○ 18-21
○ 22-25
○ 26-29
○ 30 or above



T0

Before continuing, please ask the instructor to show you the tutorial.

○ I have finished the tutorial.

reminder

R0

The following questions will be based on the SaR bot in the tutorial.

○ I understand

questions - tutorial

Q1

From the messages from SaR bot, I know how the SaR bot works.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q2

The messages from SaR bot is satisfying.

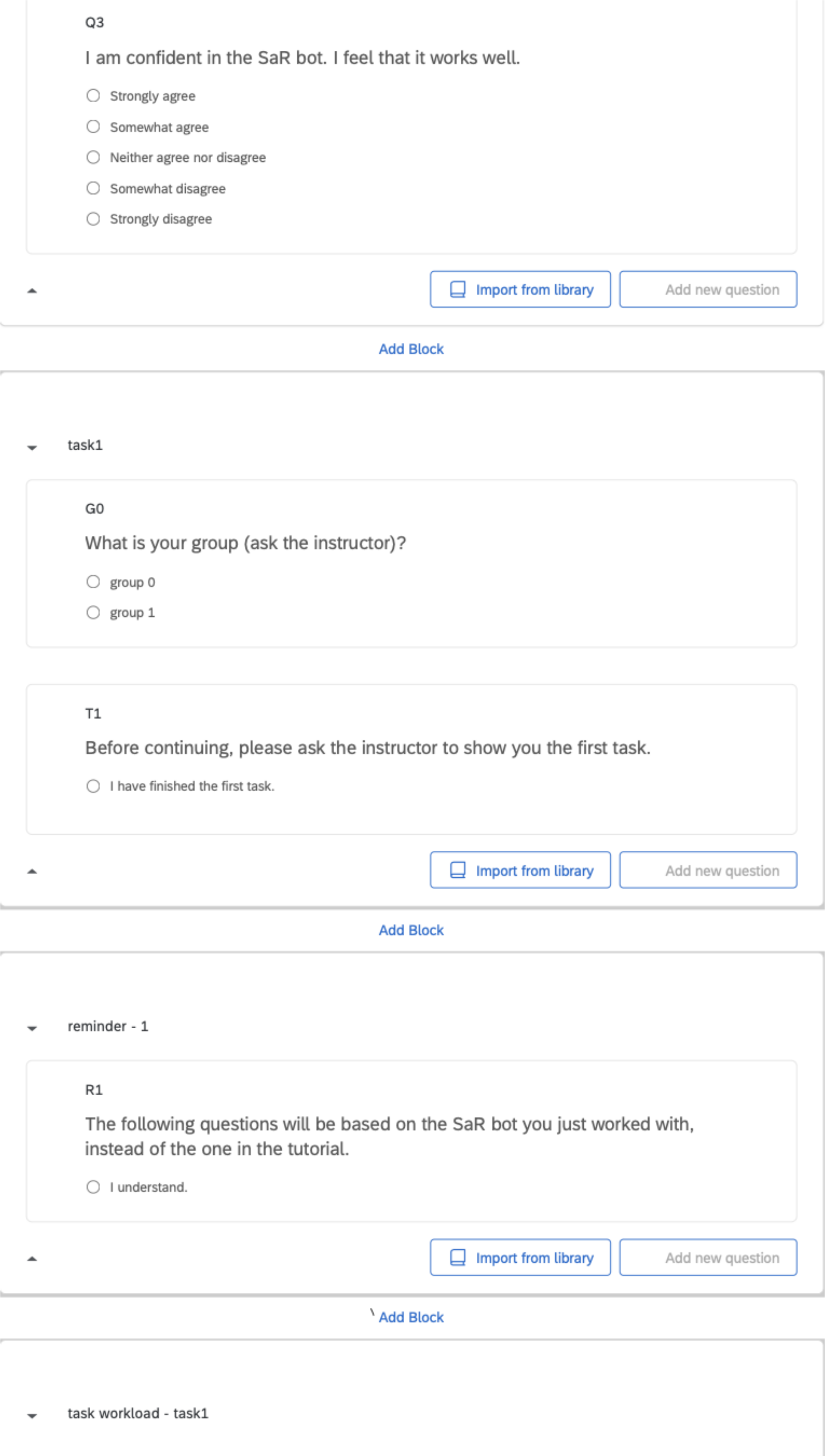
○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree



Q3

I am confident in the SaR bot. I feel that it works well.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

task1

G0

What is your group (ask the instructor)?

○ group 0

○ group 1

T1

Before continuing, please ask the instructor to show you the first task.

○ I have finished the first task.

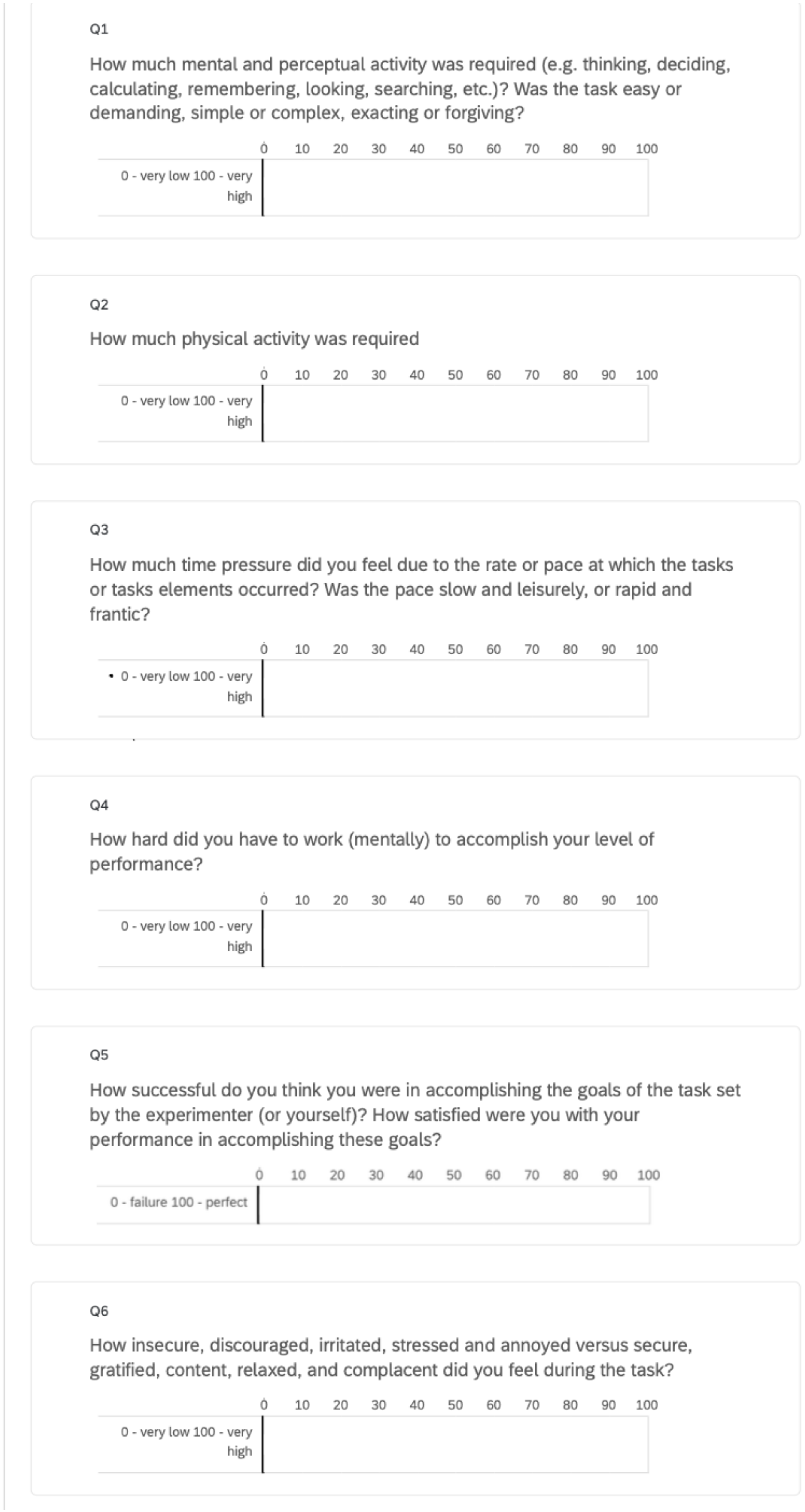
reminder - 1

R1

The following questions will be based on the SaR bot you just worked with, instead of the one in the tutorial.

○ I understand.

task workload - task1



Q1

How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

	0	10	20	30	40	50	60	70	80	90	100
0 - very low 100 - very high											

Q2

How much physical activity was required

	0	10	20	30	40	50	60	70	80	90	100
0 - very low 100 - very high											

Q3

How much time pressure did you feel due to the rate or pace at which the tasks or tasks elements occurred? Was the pace slow and leisurely, or rapid and frantic?

	0	10	20	30	40	50	60	70	80	90	100
• 0 - very low 100 - very high											

Q4

How hard did you have to work (mentally) to accomplish your level of performance?

	0	10	20	30	40	50	60	70	80	90	100
0 - very low 100 - very high											

Q5

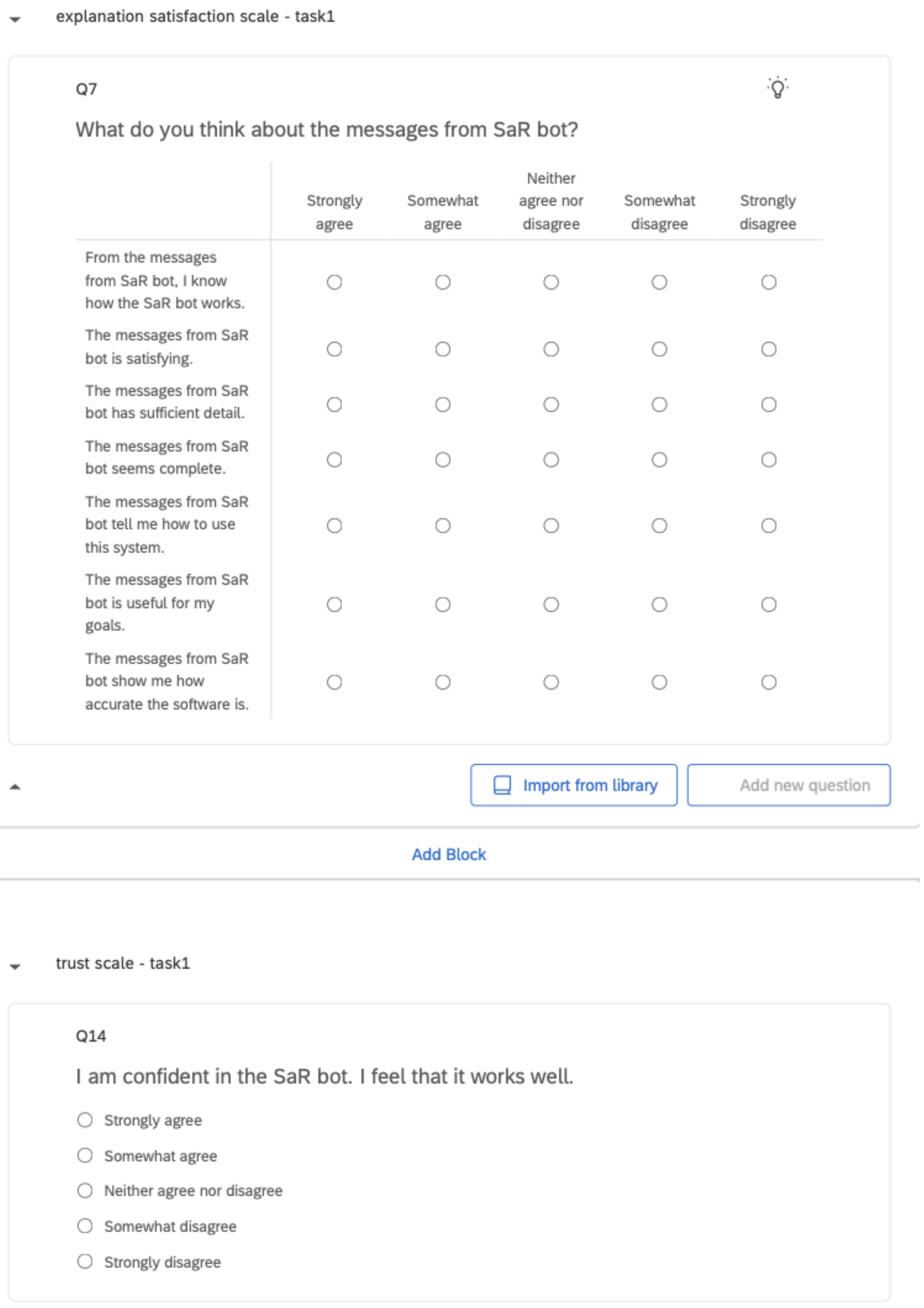
How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

	0	10	20	30	40	50	60	70	80	90	100
0 - failure 100 - perfect											

Q6

How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

	0	10	20	30	40	50	60	70	80	90	100
0 - very low 100 - very high											



explanation satisfaction scale - task1

Q7

What do you think about the messages from SaR bot?

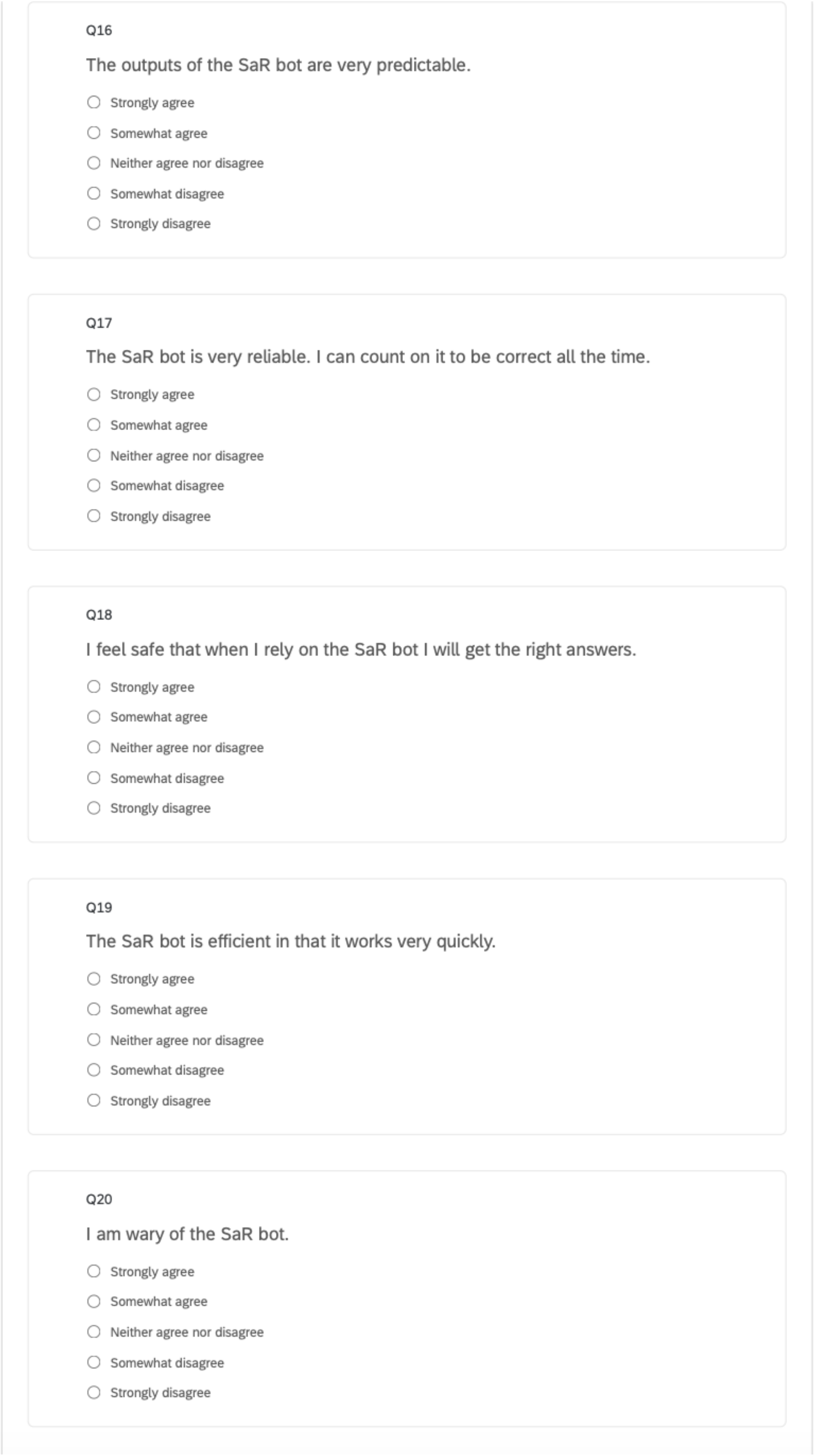
	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
From the messages from SaR bot, I know how the SaR bot works.	○	○	○	○	○
The messages from SaR bot is satisfying.	○	○	○	○	○
The messages from SaR bot has sufficient detail.	○	○	○	○	○
The messages from SaR bot seems complete.	○	○	○	○	○
The messages from SaR bot tell me how to use this system.	○	○	○	○	○
The messages from SaR bot is useful for my goals.	○	○	○	○	○
The messages from SaR bot show me how accurate the software is.	○	○	○	○	○

trust scale - task1

Q14

I am confident in the SaR bot. I feel that it works well.

- ○ Strongly agree
- ○ Somewhat agree
- ○ Neither agree nor disagree
- ○ Somewhat disagree
- ○ Strongly disagree



Q16

The outputs of the SaR bot are very predictable.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q17

The SaR bot is very reliable. I can count on it to be correct all the time.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q18

I feel safe that when I rely on the SaR bot I will get the right answers.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q19

The SaR bot is efficient in that it works very quickly.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q20

I am wary of the SaR bot.

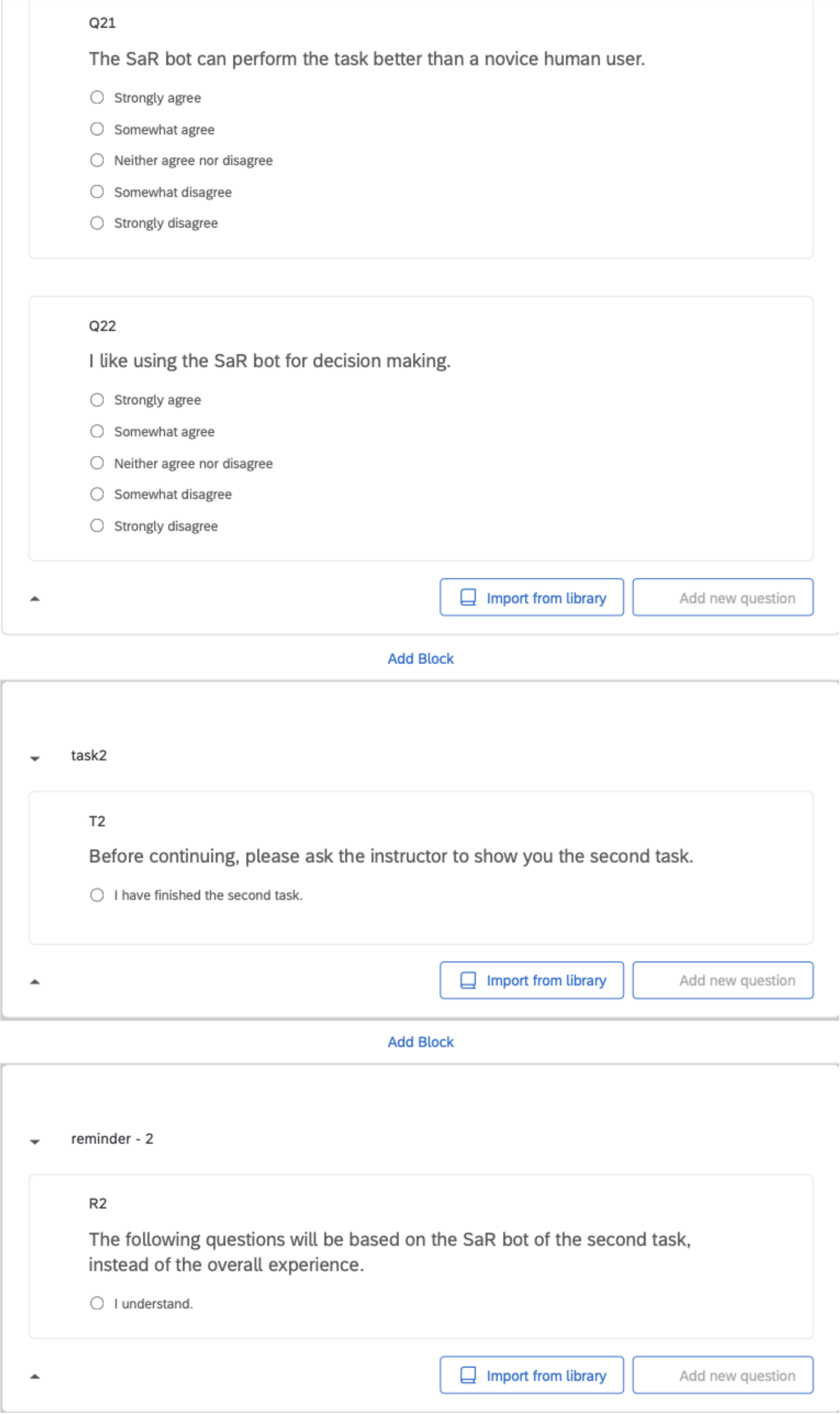
○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree



Q21

The SaR bot can perform the task better than a novice human user.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q22

I like using the SaR bot for decision making.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

task2

T2

Before continuing, please ask the instructor to show you the second task.

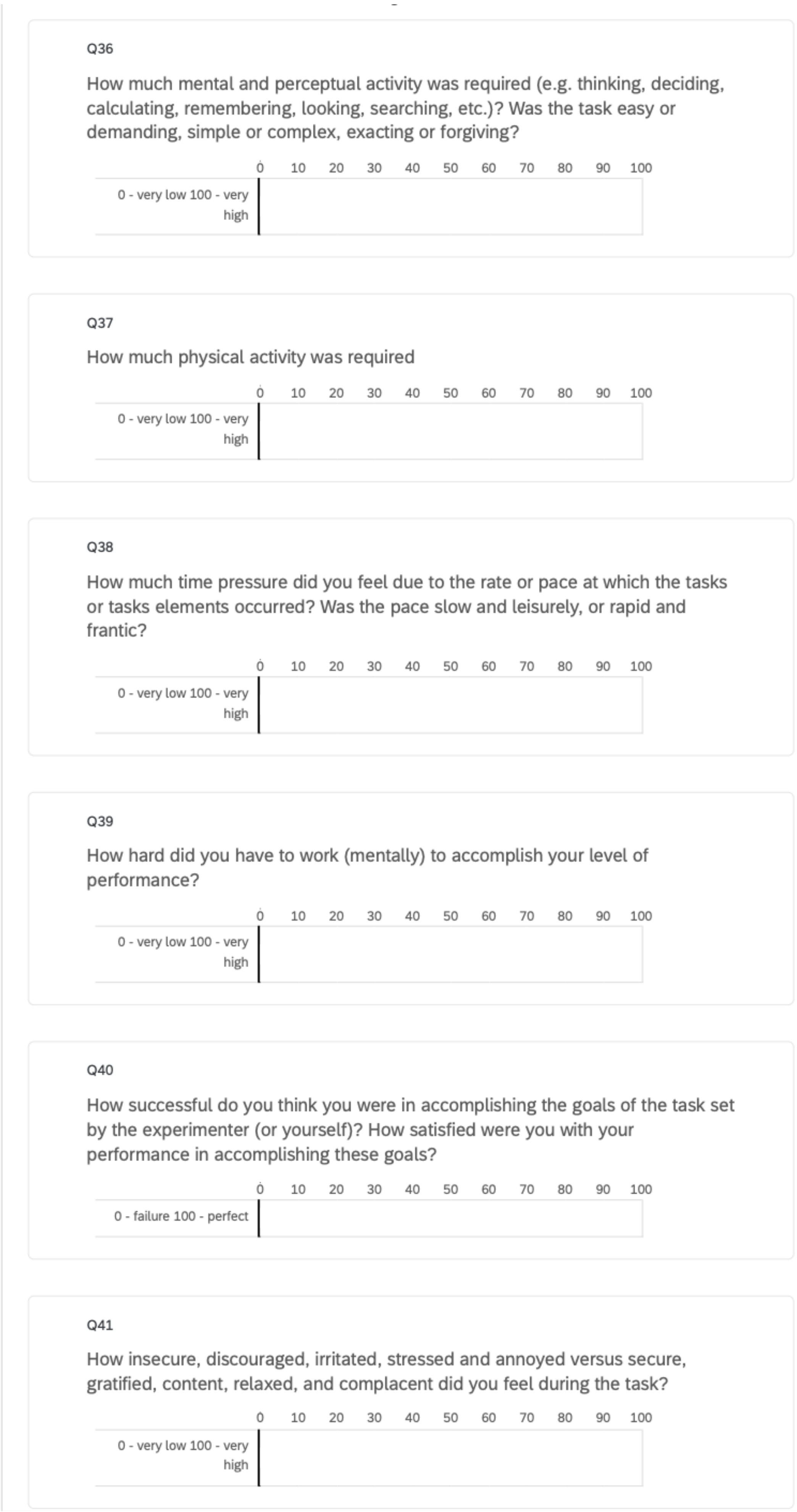
○ I have finished the second task.

reminder - 2

R2

The following questions will be based on the SaR bot of the second task, instead of the overall experience.

○ I understand.



Q36

How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

0 10 20 30 40 50 60 70 80 90 100

0 - very low 100 - very high

Q37

How much physical activity was required

0 10 20 30 40 50 60 70 80 90 100

0 - very low 100 - very high

Q38

How much time pressure did you feel due to the rate or pace at which the tasks or tasks elements occurred? Was the pace slow and leisurely, or rapid and frantic?

0 10 20 30 40 50 60 70 80 90 100

0 - very low 100 - very high

Q39

How hard did you have to work (mentally) to accomplish your level of performance?

0 10 20 30 40 50 60 70 80 90 100

0 - very low 100 - very high

Q40

How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

0 10 20 30 40 50 60 70 80 90 100

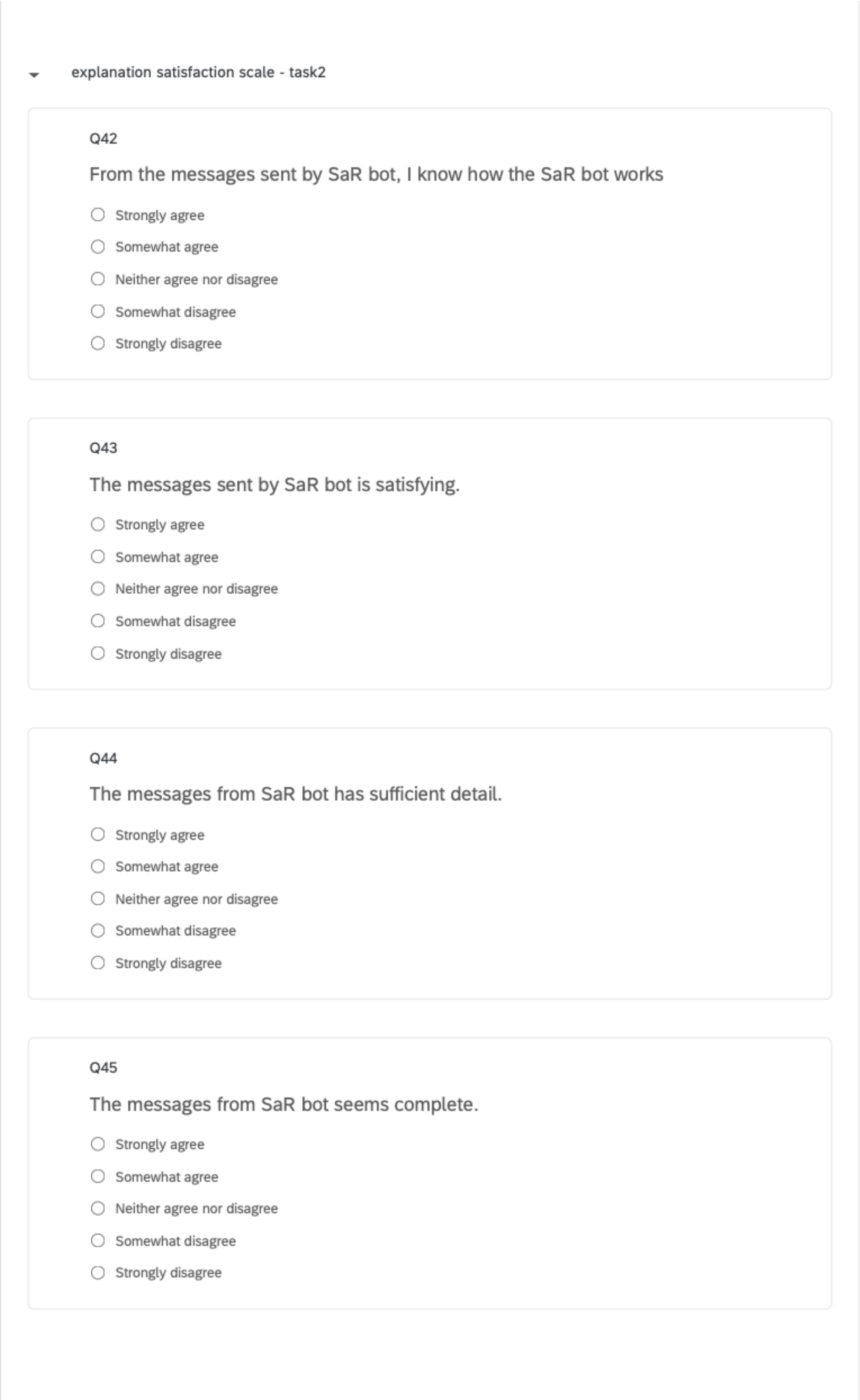
0 - failure 100 - perfect

Q41

How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

0 10 20 30 40 50 60 70 80 90 100

0 - very low 100 - very high



▾ explanation satisfaction scale - task2

Q42

From the messages sent by SaR bot, I know how the SaR bot works

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q43

The messages sent by SaR bot is satisfying.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q44

The messages from SaR bot has sufficient detail.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q45

The messages from SaR bot seems complete.

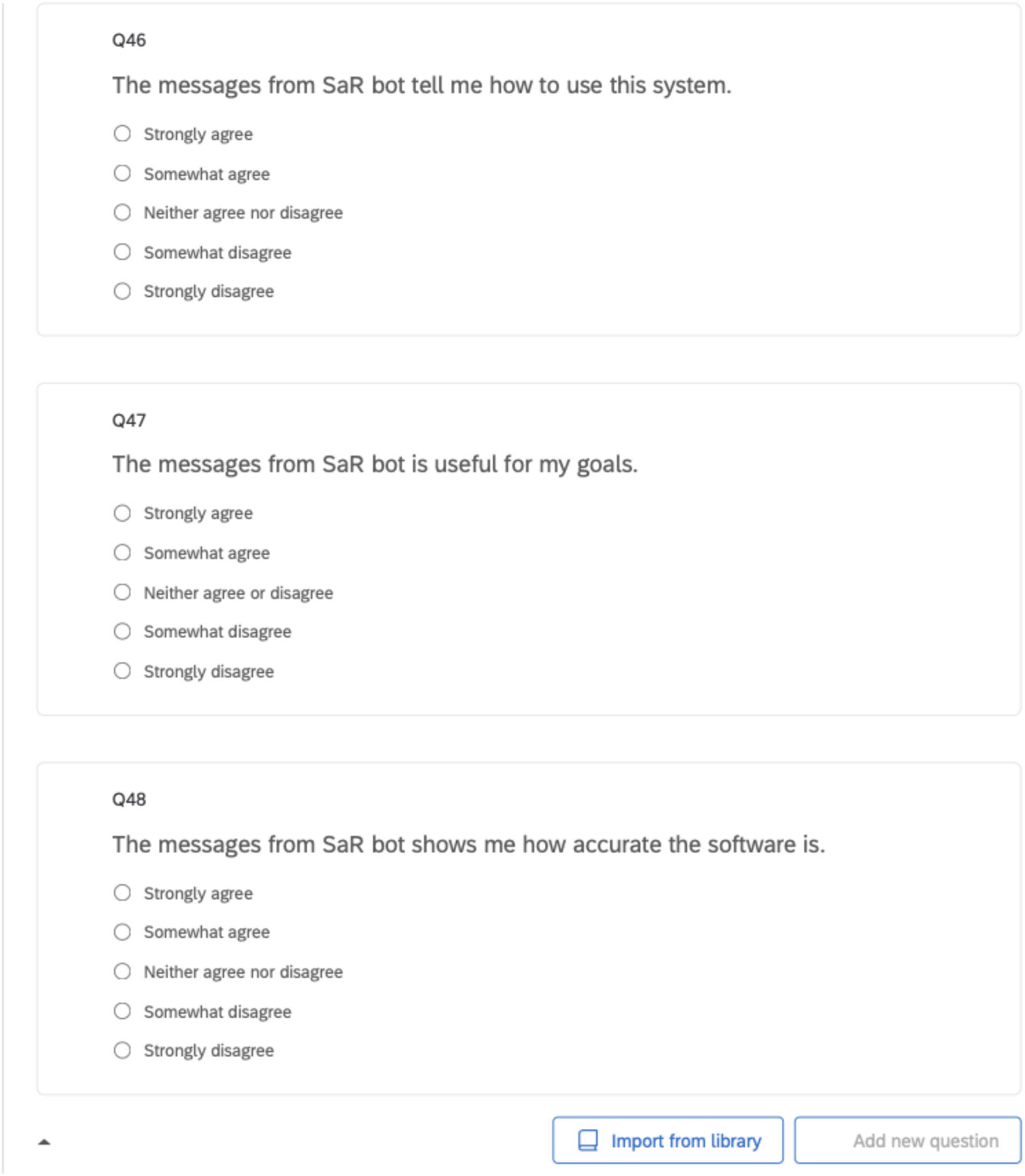
○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree



Q46

The messages from SaR bot tell me how to use this system.

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

○ Strongly disagree

Q47

The messages from SaR bot is useful for my goals.

○ Strongly agree

○ Somewhat agree

○ Neither agree or disagree

○ Somewhat disagree

○ Strongly disagree

Q48

The messages from SaR bot shows me how accurate the software is.

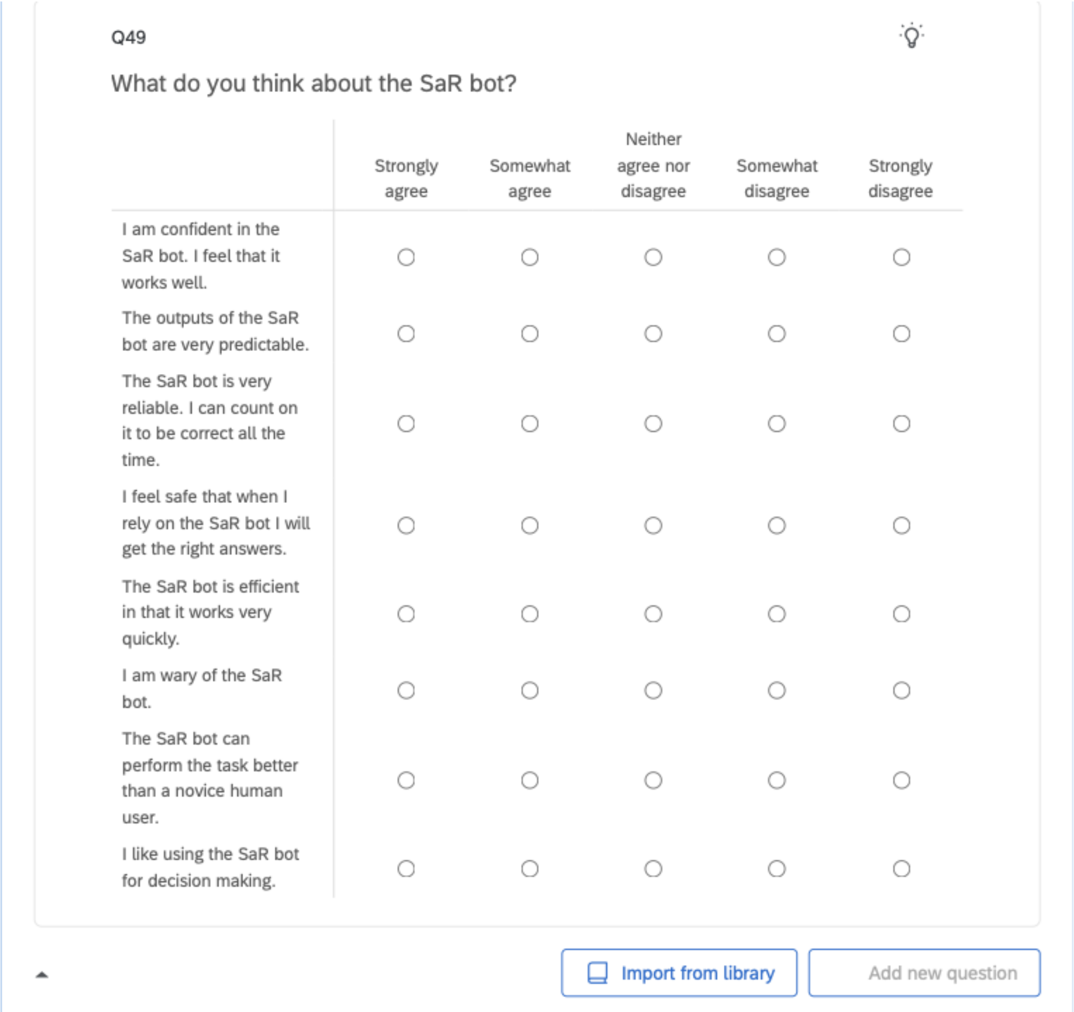
○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree

○ Somewhat disagree

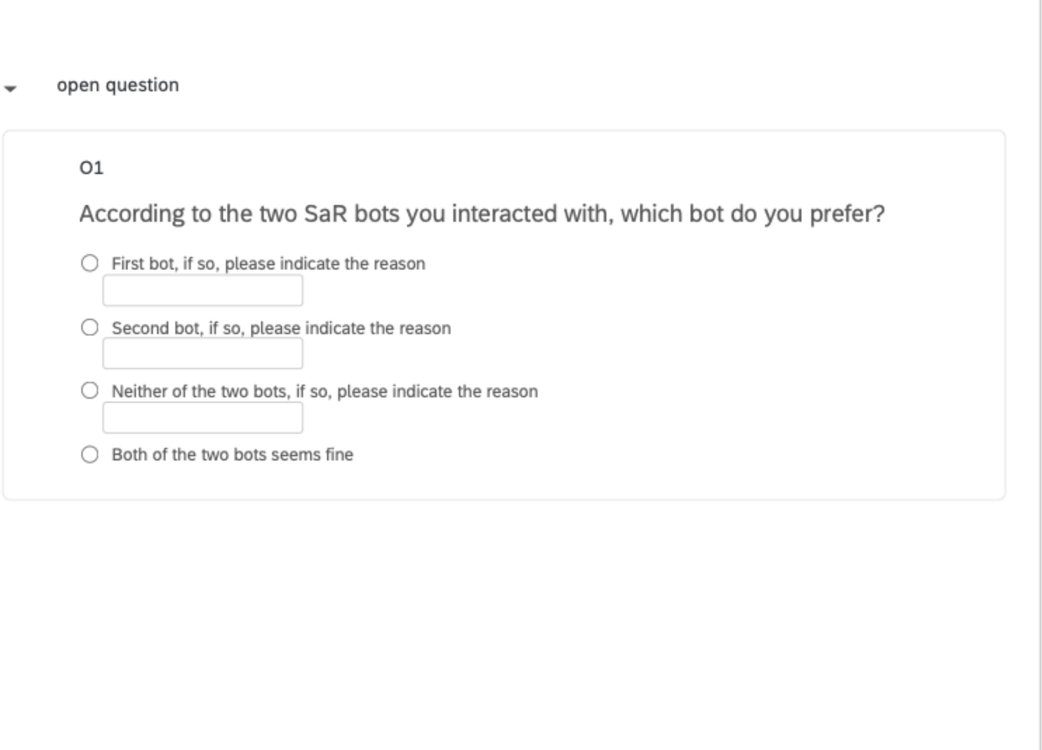
○ Strongly disagree



Q49

What do you think about the SaR bot?

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
I am confident in the SaR bot. I feel that it works well.	○	○	○	○	○
The outputs of the SaR bot are very predictable.	○	○	○	○	○
The SaR bot is very reliable. I can count on it to be correct all the time.	○	○	○	○	○
I feel safe that when I rely on the SaR bot I will get the right answers.	○	○	○	○	○
The SaR bot is efficient in that it works very quickly.	○	○	○	○	○
I am wary of the SaR bot.	○	○	○	○	○
The SaR bot can perform the task better than a novice human user.	○	○	○	○	○
I like using the SaR bot for decision making.	○	○	○	○	○



O1

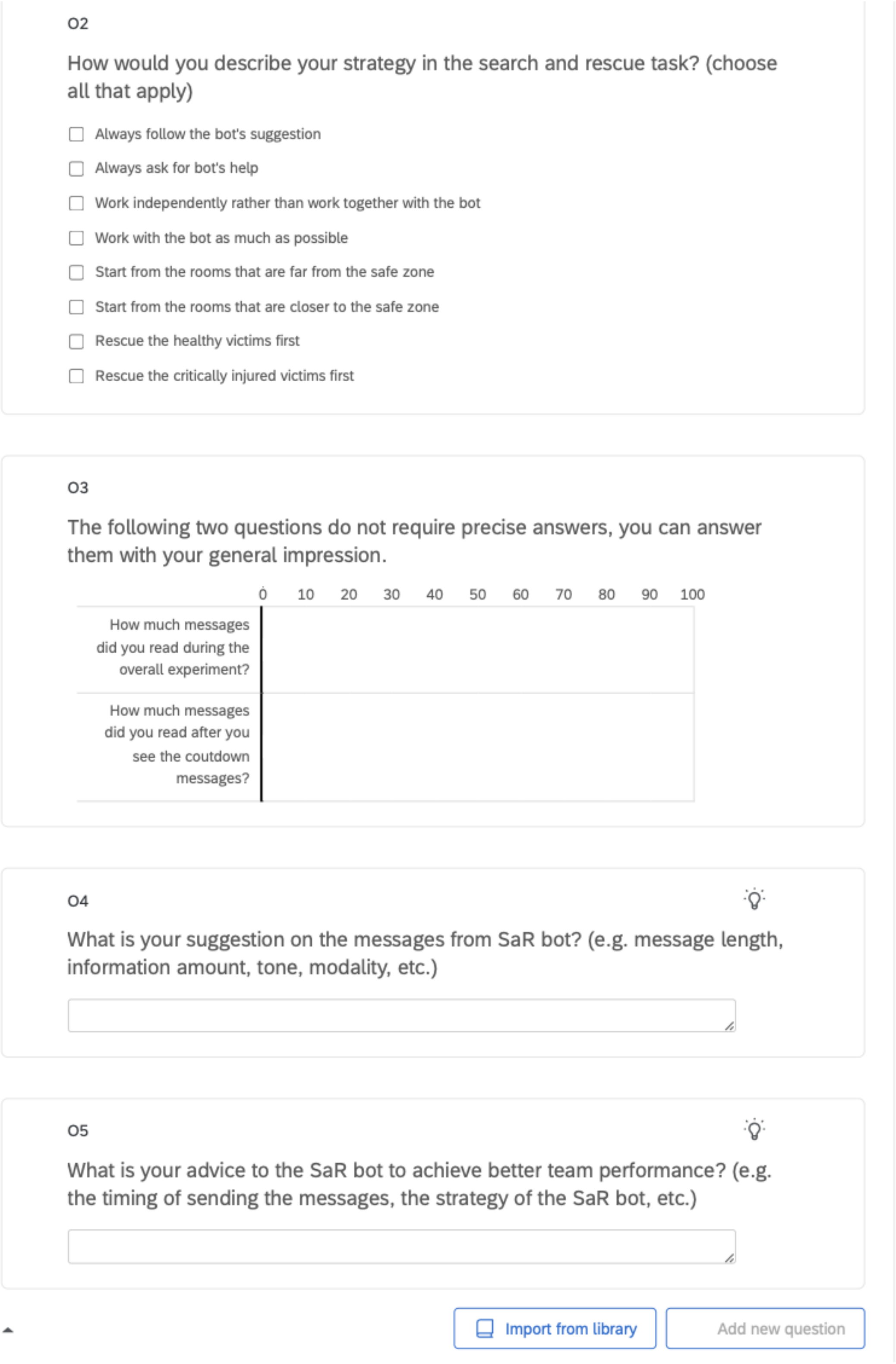
According to the two SaR bots you interacted with, which bot do you prefer?

○ First bot, if so, please indicate the reason

○ Second bot, if so, please indicate the reason

○ Neither of the two bots, if so, please indicate the reason

○ Both of the two bots seems fine



O2

How would you describe your strategy in the search and rescue task? (choose all that apply)

☐ Always follow the bot's suggestion

☐ Always ask for bot's help

☐ Work independently rather than work together with the bot

☐ Work with the bot as much as possible

☐ Start from the rooms that are far from the safe zone

☐ Start from the rooms that are closer to the safe zone

☐ Rescue the healthy victims first

☐ Rescue the critically injured victims first

O3

The following two questions do not require precise answers, you can answer them with your general impression.

	0	10	20	30	40	50	60	70	80	90	100
How much messages did you read during the overall experiment?											
How much messages did you read after you see the coutdown messages?											

O4

What is your suggestion on the messages from SaR bot? (e.g. message length, information amount, tone, modality, etc.)

O5

What is your advice to the SaR bot to achieve better team performance? (e.g. the timing of sending the messages, the strategy of the SaR bot, etc.)