Thesis

The impact of expressing emotion within explainable AI in human-agent teamwork

Sunwei Wang

Delft University of Technology

TU Delft

Thesis

The impact of expressing emotion within explainable AI in human-agent teamwork

by

Sunwei Wang

Student Name	Student Number
Sunwei Wang	4345967

Thesis advisor:	Myrthe Tielman
Thesis committee member:	Mark Neerincx
External thesis committee member:	Jie Yang
Daily co-supervisor:	Ruben Verhagen
Project Duration:	July, 2023 - March, 2024
Faculty:	Faculty of Computer Science, Delft

TU Delft

Acknowledgement

Completing this thesis marks the end of my Master's program, a journey filled with both educational and personal growth. I owe a huge thank you to everyone who has helped me along the way.

Firstly, I'm grateful to Myrthe Tielman for being my thesis advisor and guiding me through the process, and to Ruben Verhagen for his continuous support and feedback on my thesis. Our weekly meetings were a great space to share and discuss ideas. Mark Neerincx, thank you for your valuable advice and feedback during both my first stage and the green light meeting. Thanks to Jie Yang for being part of the thesis committee.

A big thank you goes to all the friends and family who took time out for my experiment, making this project possible. Special thanks to my wife for always being there, and to my parents and host parents for their endless support and encouragement.

Sunwei Wang
Delft, March 2024

Abstract

With the increasing development of artificial intelligence (AI), there is a more significant opportunity for humans and agents to collaborate in teamwork. In Human-Agent Teamwork (HAT) settings, collaboration requires communication, and the agent displaying emotion can impact how human teammates communicate and work together with the agent. This study investigated the impact of an explainable agent expressing emotion within explanations in a teamwork setting. We investigated how integrating an emotional component into an agent's explanations influences trust in the agent, as well as humans' perceptions of the agent's anthropomorphism, animacy, likeability, and overall team performance when collaborating with the agent. With this goal, a pre-study was conducted using a focus-group meeting to investigate the relevant emotions to display in a simulated Search and Rescue (SAR) task and how these emotions can be incorporated into Explainable AI (XAI). Next, we conducted an in-between subject-controlled experiment to study the effects of emotional components in explanations. The participants were divided into experimental and control groups, collaborating with agents that either displayed emotion or no emotion. The participants had to carry out a SAR task where they worked together with the agent to rescue victims. Our results confirmed that an agent displaying emotions increased perceived likeability, animacy, and anthropomorphism. Among these three, likeability and animacy are positively associated with trust. In contrast, an increase in anthropomorphism is associated with a decrease in trust. From the results, we could not conclude that team performance is directly affected by having emotion in the explanation. However, the results showed that emotion increases the messages sent from the human to the agent, and this increase in communication led to higher team performance.

Keywords human-agent teamwork, explainable AI, user study, communication, emotion, anthropomorphism

Contents

1

Introduction

As robotics and artificial intelligence (AI) continue to advance, the integration of robots into various aspects of human life becomes increasingly prevalent. One significant area of interest is the collaboration between humans and agents in teamwork scenarios. For instance, in healthcare settings, robots and medical staff work together to enhance patient care [Beasley et al., 2012]. In education, AI can assist interactive learning, aiding teachers and engaging students in new ways [Holstein et al., 2019]. Effective human-agent teamwork demands seamless interaction and communication [Anjomshoae et al., 2019]. Emotion has significant relevance in human teamwork dynamics. However, there has been limited exploration of its role in human-robot teamwork. This thesis aims to explore the effects of emotional expression in human-agent teamwork and understand how it influences team dynamics and overall performance.

In our everyday interactions, emotions play a vital role in helping us connect and understand each other [Döring, 2003]. Consider a time when a friend's warm smile and encouraging words helped you during a tough moment, or how a colleague's excitement inspired you to do your best. These emotional signals act as the glue that holds our relationships together, making us trust each other more and work better as a team [Nair et al., 2005].

As we enter the field of AI and agent technology, there is also an opportunity to enable machines to express emotions [Adadi and Berrada, 2018]. Imagine a robot working in a team that not only performs its job well but also shares how it feels – whether it is happy with the progress, concerned about something, or just excited to work together. This situation is similar to having a team member who can express encouragement with a simple nod or offer comfort through words, contributing to stronger team dynamics.

Think about a future where emergency teams include humans and robots working side by side. In critical situations, having robots that can show emotions might be helpful for their human teammates. A robot expressing urgency or worry could help the whole team stay focused and avoid chaos. It is as if having an agent who understands the importance of the moment and guides everyone with a clear sense of purpose.

Investigating how agents express emotions in teamwork is not just a matter of science. It is a journey to make our interactions with technology more human and

meaningful. By studying how emotional cues work in robot explanations [Anjomshoae et al., 2019], we are trying to uncover the subtle things that make teamwork successful. In doing so, we are shaping a future where humans and agents work together seamlessly, making the most out of what each brings to the table.

However, ”Anthropomorphism can lead to an inaccurate understanding of biological processes in the natural world,” says Patricia Ganea, a psychologist at Toronto University [Epley et al., 2007]. While robots can mimic emotions, they do not genuinely experience them like humans do. This could lead to misunderstandings or false expectations, where humans might perceive a robot's emotions as genuine, potentially affecting how they interpret and respond to the robot's actions. This raises a question in human-robot interaction research: How can robots convey emotions in their explanations to human teammates in a way that benefits teamwork? In this study, we examine the mechanisms and techniques that enable robots to effectively communicate emotions, ensuring that human interactions with robots in critical contexts, such as search and rescue missions, are both efficient and emotionally supportive.

The real challenge lies in understanding how much we can trust the emotions displayed in robots' explanations. We seek to explore the positive aspects of having robots show emotions in their explanations while being mindful not to expect them to react as humans do. This balancing act between the benefits and potential challenges of robots showing emotions in their explanations makes our study even more thought-provoking. It helps us see the bigger picture and ensures our research is relevant in the evolving landscape of human-agent collaboration.

1.1. Motivation

The integration of robots into teamwork environments has far-reaching implications for both scientific and societal contexts. For instance, robots can assist medical teams with surgeries in healthcare, improving precision and reducing fatigue [Beasley et al., 2012]. In manufacturing, collaborative robots can work alongside humans to increase productivity and safety [Sherwani et al., 2020]. In emergency response, robots can perform tasks in hazardous environments, reducing risk for human first responders [Stormont, 2005]. Understanding how emotional expression by robots influences human-robot interaction is vital for designing efficient and harmonious team dynamics. Additionally, this research addresses several motivations:

1. Enhancing Human-Robot Collaboration: By investigating the effects of emotional expression in human-robot teamwork, it will assist us in designing robots that could better display human emotions in conversation with human partners, leading to more effective and intuitive collaboration.
2. Human-Robot Trust and Acceptance: Emotions play a significant role in building trust and rapport among human team members. Studying emotional expression in robots helps determine if humans are more likely to trust and accept robots as valuable team members.
3. Psychological Impact on Humans: Robots that display emotions could evoke emotional responses in humans. Examining these emotional responses helps us understand the psychological impact of human-robot interaction, particularly when emotions are involved.

4. Ethical Considerations: As robots become more emotionally expressive, ethical considerations arise concerning how these robots may influence human emotions and decision-making. This research will contribute to the ongoing discussion on the responsible deployment of emotionally expressive robots.

This thesis seeks to contribute to the field of human-agent interaction by shedding light on the effects of robotic emotional expression in human-agent teamwork. By studying the impact of robotic emotional expression on collaboration, trust, and team performance, this research aims to inform the development and deployment of emotionally intelligent agents in various teamwork scenarios, aiming to improve human-agent collaboration and enhance the user experience.

1.2. Research Question

Main RQ:

- How does the expression of emotion in the explanations of an agent impact human-agent interaction and teamwork?

Sub RQs:

- What are the relevant emotions for an agent to express in teamwork settings?
- How could we incorporate emotions into an agent's explanation?
- What are the dependent variables that could be influenced by having emotions in the agent's explanation?

Our study of the RQs focuses on uncovering the effect of these emotional expressions from agents on how we interact with them and how we collaborate in teams. We want to know if the emotions that agents show in their explanations impact our teamwork. It is like figuring out whether an agent's smile or concern can make us work together even better.

We will also explore whether people believe that these agent emotions are genuine and if they feel like the agents can truly understand their own feelings. This is related to the subject of anthropomorphism [Złotowski et al., 2015]. Studying anthropomorphism is important because it affects how much we trust and connect with these agents as part of our team.

There are exciting possibilities like agents helping us feel more motivated or reassured during challenging tasks. However, there could also be challenges, such as misunderstandings or discomfort caused by an agent expressing emotions. We want to weigh these pros and cons to see if emotional agents can truly enhance team collaboration.

In our investigation, we will investigate the range of emotions that artificial intelligence can express during teamwork situations. Based on existing work of how others accurately measure the emotions in their research, such as from [Gratch and Marsella, 2004], we will also conduct a pre-study to help us narrow down the suitable emotions that we should use for our design. This way, we can gather reliable insights into how robot emotions affect our interactions and teamwork dynamics.

In summary, our study seeks to solve the impact of emotional expressions from agents on our interactions and teamwork. By understanding the potential benefits

and possible drawbacks, we hope to contribute to the future development of agents that can truly connect with their human team members and make our collaborative efforts even stronger.

2

Background & Related work

This chapter examines the foundational background and its related studies, focusing on three key elements that are crucial to my research questions: Explainable Artificial Intelligence (XAI), Human-Agent Teamwork (HAT), and Emotion.

2.1. Explainable AI (XAI)

XAI refers to developing artificial intelligence systems whose actions, recommendations, and underlying decision-making processes are comprehensible and transparent to human users [Langley et al., 2017]. This transparency is essential in different domains, including healthcare, finance, and, more prominently, human-agent collaboration [Vilone and Longo, 2021].

The field of XAI contains a broad range of approaches to make AI systems more understandable to humans. Within this field, two primary categories emerge based on the focus and nature of the AI systems being explained: data-driven XAI and goal-driven XAI Verhagen et al. [2021].

2.1.1. Data-driven XAI

In recent years, the field of Explainable AI (XAI) has gained significant traction, with a substantial focus on making machine/deep learning models' decisions understandable and transparent to users [Vilone and Longo, 2021]. According to [Anjomshoae et al., 2019], data-driven XAI is about understanding a decision of a "black-box" machine learning algorithm given the data as an input. This branch of XAI aims to explore how these models make predictions or decisions based on the input data they are fed. The challenge here is to interpret the rationale behind the predictions of the deep learning models. For example, feature attributions in image recognition tasks emphasize which parts of an image were most significant in the model's decision, providing users with direct insight into the model's thought process [Baumgartner et al., 2018]. Another concrete example of an explanation type in data driven XAI would be the contrastive explanation, which is an explanation that compares the agent's output to an alternative counterfactual output [Neerincx et al., 2018].

2.1.2. Goal-driven XAI

Goal-driven XAI is about an explainable agency, which involves AI explaining the actions and reasons leading to their decisions. Goal-driven XAI includes autonomous agents and robots designed to operate and make decisions on their own within their surroundings to achieve specific goals, whether assigned or self-determined [Biran and Cotton, 2017]. Goal-driven XAI is a field focused on developing robots or agents capable of explaining their actions in a way that is understandable to the everyday user. According to [Sado et al., 2023], "The explanations would assist the user in creating a Theory of Mind (ToM), comprehending the agent's behavior, and contribute to greater collaboration between the user and the agent." Another example includes belief-based explanations, which clarify the AI's goals and intentions, contributing to improved understanding and teamwork in human-agent interactions [Rao et al., 1995].

XAI's relevance spans various domains, particularly in enhancing human-agent collaboration. For instance, in disaster response training, AI agents that provide rational explanations for their actions can significantly improve trainees' decision-making skills and overall training effectiveness [Core et al., 2006, Graesser et al., 2005]. Both data-driven and goal-driven approaches are essential to help promoting the field of XAI and ensuring that AI systems are transparent, trustworthy, and aligned with human values and expectations. We will explore the topic of emotion in explainable AI and focus on the area of explanation phases.

2.1.3. Explanation phase

To comprehensively investigate the impact of AI expressing emotions within explanations in the context of human-robot teamwork, it is essential to establish a structured framework for the explanation process. This framework encompasses three crucial explanation phases, as stated by prior research [Neerincx et al., 2018]. In our explanation generation design chapter, we will discuss the generation and communication phases in Section 3.3. In this section, we will go into detail about the related works of the three explanation phases.

The first phase, explanation generation, is dedicated to crafting justifications for the actions or results achieved by the AI agent. It involves the "why" behind an AI's decision or behavior. The implementation of this phase is influenced by the AI model employed by the agent, such as a Belief-Desire-Intention (BDI) agent [Rao et al., 1995]. This phase draws upon diverse sources, including the agent's goals [Broekens et al., 2010], desires [Kaptein et al., 2017a], and even its emotions, as outlined in the existing literature [Kaptein et al., 2017b].

The second and the third explanation phases are also full of value; they are the explanation communication and explanation reception. [Lewis, 2020] described an explanation as a set of information clarifying the causes behind events, which is particularly valuable in the context of Human-Agent Teamwork (HAT). Such information is vital to enhance coordination within human-agent teams. [Neerincx et al., 2018] pointed out that in the phase of explanation communication, both the form and content of the explanation needs to be considered. Explanations can take multiple forms—ranging from texts and audio to visuals like images and videos [Oei and Patterson, 2013]. Moreover, the content of these explanations can be adapted to the specific context of the scenario or customized to meet the preferences of the individual user [Anjomshoae

et al., 2019].

Moving to the third phase, explanation reception, the focus is on how effectively humans comprehend the given explanation. Existing studies on Explainable AI (XAI) reception have been conducted (e.g., [Narayanan et al., 2018]). However, there is a gap in empirical research involving actual human task performers needing explanations in human-agent settings (as indicated in [Miller, 2019]). Our research aims to bridge this gap by assessing how participants respond to the introduction of emotions in explanations. This experimental investigation forms the core of our research question, as we seek to understand participants' reactions to these emotional additions, which constitutes the primary objective of our study.

Furthermore, to understand how emotions can be effectively integrated into the explanation process, this research considers the selection of functional feelings and emotions. As discussed by [Nair et al., 2005], emotions are relevant in human teamwork dynamics. However, their role in human-robot teamwork contexts needs to be explored more. This study, therefore, uses the framework of these explanation phases, coupled with the strategic inclusion of emotions, investigating influence on the interactions and outcomes of human-robot teamwork, shedding light on an emerging and vital facet of AI research in the evolving landscape of artificial intelligence.

2.2. Human-Agent Teamwork (HAT)

The synergy between human intelligence and AI in HAT settings aims to achieve better outcomes beyond an individual's capabilities. Researches such as [Caldwell et al., 2022, Wang et al., 2019] have been conducted to show that agents are treated as teammates rather than tools in multiple contexts of collaboration between humans and agents. Agents can take the responsibility of a teammate in HAT settings, and contribute to shared team goals [Zhang et al., 2024]. As [Chen et al., 2018] stated in their research, over the years, it has become apparent that sharing more reasoning details from an agent can enhance trust and performance. Yet, it's crucial to avoid overwhelming human teammates. Recently, the demand for tailored agent explanations has grown, highlighting the importance of adjusting the amount of shared reasoning to suit both the user's needs and the context. Effective collaboration and coordination, supported by mutual trust and clear communication such as providing explanations, are essential for success in various fields, including medical and firefighting domains [Salas et al., 1997, Teaming, 2022].

In the disaster response situation, let us consider a scenario where a group of trainees is practicing emergency response in a simulated earthquake. Emotionally intelligent AI agents are integrated into the team. These AI agents can explain the rationale behind each step of the emergency response. For instance, when rescuers are instructed to prioritize searching for survivors over securing the area, the AI agent can explain, "In a real disaster, saving lives is the top priority because immediate medical attention can be life-saving. We secure the area after ensuring everyone's safety." This explanation helps rescuers understand the reasoning behind their actions and learn the critical decision-making process in a disaster scenario [Core et al., 2006].

Another example is a virtual mathematics study team. The AI agent within the system engages in a natural dialog with the student, explaining various mathematical principles. For example, the AI agent might say, "Let us break down this equation step

by step when introducing a challenging equation. By factoring these terms, we can simplify the problem and make it more manageable. This approach helps increase the understanding of the underlying concepts and solve similar equations in the future." Through this dialog and explanation, the student gains a deeper understanding of the subject, improving the overall effectiveness in solving the team mathematics problems [Graesser et al., 2005].

In both scenarios, the explanations provided by AI agents enhance the experience by making the rationale behind actions or concepts more transparent, facilitating better comprehension, and ultimately increasing the effectiveness of the performance of the team. Understanding AI's decisions and actions is crucial in ensuring trust, accountability, and effective teamwork between humans and robots [Anjomshoae et al., 2019].

2.2.1. XAI in HAT

The integration of XAI within HAT highlights the necessity of providing clear and understandable explanations for AI actions. Despite existing research on different explanation types, there remains a significant gap in exploring emotional aspects of XAI, which is critical for fostering deeper human-agent collaboration [Madsen and Gregor, 2000, Johnson and Vera, 2019].

There are additional works that have been contributed to the field. [Verhagen et al., 2021] have presented a two-dimensional framework that defines and relates these concepts concisely and coherently, yielding a classification of three types of AI systems: incomprehensible, interpretable, and understandable.

In the collaborative landscape of Human-Agent Teamwork (HAT), optimal performance depends on the expertise of each team member to achieve better outcomes [Salas et al., 1997]. The synergy between human intelligence and artificial intelligence is envisioned to elevate performance beyond the capabilities of either entity individually [Teaming, 2022]. To realize this synergy, effective collaboration and coordination become imperative elements within HAT [Stowers et al., 2021]. Successful collaboration, as established in the literature [Johnson and Vera, 2019, Schoonderwoerd et al., 2022, Harbers et al., 2011a,b], is dependent on information sharing, and transparent and reasoned explanations.

Critical to the success of HAT communication are factors such as transparency, mutual trust, understandability, and explainability [Madsen and Gregor, 2000, Salas et al., 1997, Harbers et al., 2011a]. The understanding of decisions made by the agent relies heavily on providing explanations and offering humans insight into the inner workings of the agent. Unfortunately, modern HAT systems often fall short in delivering these important factors, particularly in terms of poor explainability, leading to decreased collaborative performance [Madsen and Gregor, 2000, Johnson and Vera, 2019].

XAI research aims to enhance AI-human interactions, increase trust, and improve team dynamics. Challenges include modeling a wide range of emotions, integrating them with explanations, and ensuring human perception of authenticity. Addressing these challenges requires interdisciplinary efforts to create emotional AI agents for effective teamwork. Exploring emotional XAI in HAT settings could improve AI systems' functionality and acceptance, offering more profound insights into human-AI emotional

dynamics [Verhagen et al., 2021, Anjomshoae et al., 2019, Neerincx et al., 2018].

2.3. Emotion in Collaboration

Emotions play a pivotal role in human communication and decision-making, influencing the dynamics of both human-human and human-agent interactions. The ability of AI agents to express emotions can enhance trust and likability, making agents appear more human-like and relatable [Fong et al., 2003, Nair et al., 2005].

As recognized by [Nair et al., 2005], emotions play a significant role in shaping human teamwork dynamics, yet their integration into human-robot teamwork contexts remains under-explored. Therefore, in our research, we will look closely into the integration of emotion into human-robot teamwork. Expressing emotions in AI poses unique challenges, such as the risk of anthropomorphism leading to misinterpretation of the robot's displayed emotions. This highlights the need for careful design in conveying emotions, ensuring they are appropriately matched to the context and user expectations [Bartneck et al., 2009, Zhou and Tian, 2020].

Integrating emotions into AI explanations offers a promising route to improve the efficacy of human-agent teamwork. However, balancing the benefits of emotional expressions with the potential for misunderstanding requires further research to optimize this integration for various teamwork contexts [Jiang et al., 2007, Oei and Patterson, 2013].

Emotions are pivotal in human communication, decision-making, and cognitive processes [Fong et al., 2003]. In the current landscape of artificial intelligence research, a recognized focus is directed towards understanding human emotions and their perception of artificial emotions expressed by intelligent agents, contributing to improved agent autonomy and socially acceptable interaction design. While prior works, such as the Emotional Belief-Desire-Intention agent model [Jiang et al., 2007], explored the influence of artificial emotions on decision-making within single-agent or Multi-Agent System (MAS) settings, the collaborative dynamics between humans and intelligent agents require deeper exploration.

The work of Zhou and Tian [Zhou and Tian, 2020] investigates the impact of robots' emotional expressions on collaboration outcomes and human perceptions. The emotional robots exhibited artificial emotions in a controlled experiment involving a human-robot team comprising one human and two Cozmo robots engaged in a collaborative game. In contrast, the non-emotional robots remained devoid of emotional expressions. Their research showed that non-verbal emotional expressions were practical for robots to ask for help from human teammates. Additionally, human teammates had a more pleasant experience interacting with emotional robots and perceived them as more competent. These findings provided insights into human perceptions of emotionally expressive robots and their responses to diverse robot-robot and robot-human communication designs. However, they have only conducted their experiment with 8 participants per condition for between-subjects studies. Thus, there is not enough evidence to support the results. There is another research conducted by [Fadhil et al., 2018], which also discovered that emojis can enhance enjoyment, attitude, and confidence when interacting with the conversational agent. However, they were measuring the concept of having emoticons present, not the concept of emotions. Therefore, our study would try to conduct between-subject studies with enough participants and in-

vestigate various aspects of team dynamics and team performance. One of the key aspects we will look into is the challenge for robots expressing emotions, such as anthropomorphism.

2.3.1. Anthropomorphism

Anthropomorphism is a phenomenon that describes the human tendency to see human-like shapes in the environment, as phrased by [Złotowski et al., 2015]. Anthropomorphism, in the context of robotics, refers to the attribution of human-like characteristics, qualities, or behaviors to robots. It influences both the design of the robots and how humans interact with them [Nicolas and Agnieszka, 2021]. In robotics, while machines can mimic emotions, they cannot genuinely experience them as humans do. This creates a challenge: humans may mistakenly perceive a robot's displayed emotions as authentic, affecting their interpretation and response to the robot's actions. An important question in human-robot interaction research is: How can we improve how robots convey emotions in their explanations to human teammates?

One of the fundamental researches about anthropomorphism in human-robot interaction(HRI) has been conducted by [Bartneck et al., 2009]. The authors have done a literature review on five key elements in HRI: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety, and proposed the Godspeed Questionnaire Series (GQS), which is one of the most highly cited and used questionnaire in the field of HRI. We will also utilize part of the GQS to facilitate the implementation of our user study. Looking closely at how anthropomorphism and emotion expression are connected, helps us understand why this link is so important. Anthropomorphism is not t just about making robots look or act human; it is also about how these robots display emotions. This is important because the way robots show emotions can affect how humans perceive them, how much we trust them, and how well we can collaborate with them. When we design robots to seem more human, our goal is to make interactions with them feel more natural and easy, especially when it comes to sharing emotions. This can help everyone get along better and work more effectively as a team. However, it could also cause discomfort and distrust when the emotions are not properly displayed, we should find the right balance to improve HAT.

This chapter sets the stage for investigating how emotional expressions in AI agents impact human-agent teamwork, particularly within search and rescue scenarios. The discussion underscores the importance of XAI, HAT, and emotion in developing effective and collaborative human-agent systems. The insights gathered here inform the pre-study, which explores specific emotional expressions relevant to teamwork, leading to the research question and subsequent investigation.

3

Emotional Explanation Design

3.1. Overview

In order to design the emotional explanations of the agents for our user study, we first conducted a pre-study investigating whether the selected visual and textual emotion components were appropriate or would grant the expected reaction from our participants. This pre-study was conducted in the form of a focus-group meeting, and aimed to explore the possible relevant emotions for an agent to express in a search and rescue teamwork setting and how we can incorporate these emotions into agents' explanations. By analyzing non-expert participants' reactions to various AI-expressed emotions and explanations, the study seeks to shed light on the role of emotions in enhancing human-AI collaboration.

3.2. Design

The study employed a qualitative approach, using focus group discussions to gather insights. Various scenarios where an AI agent expressed emotions were presented to participants. These scenarios mimicked potential real-world interactions in search and rescue team environments, allowing for the collection of data on instinctive human responses to AI's emotional cues.

Participants included 3 non-experts, selected from personal acquaintances such as family and friends. This criterion was chosen to ensure a diverse range of intuitive responses, as these individuals had no formal background in AI or computer science. Their lack of technical bias made their feedback especially valuable in understanding the general public perception of the design of emotion in explanations from AI.

3.2.1. Emotion design

Research on emotions and feelings has yielded various theories regarding the basic emotions, with scholars proposing differing sets based on their findings. [Cacioppo et al., 1993] identified four fundamental emotions: fear, anger, joy, and sadness. Expanding upon this, [Ekman et al., 1999] listed six basic emotions, namely anger, disgust, happiness, sadness, fear, and surprise. [Plutchik, 2001] proposed an even broader spectrum, suggesting eight basic emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. Additionally, [Gu et al., 2019] introduced a set

of seven universal emotions: happiness, surprise, contempt, sadness, fear, disgust, and anger. These variations highlight the diversity of perspectives within the field and reflect the ongoing debate and exploration, with some studies introducing new dimensions or combinations to these foundational emotional categories, for example, Figure 3.1.

Most of the existing emotion studies have been done on humanoid robots, but there is a study " Measuring emotions of robot operators in urban search and rescue missions" by [Mioch et al., 2013], which looked into the question: what kind of emotions do firefighters show during Urban Search And Rescue (USAR) missions? The answer was that firefighters reported that they did not experience emotions during the execution of the scenario. However, we need to investigate whether this is due to the high-stress situation that firefighters are in, or the delayed emotional responses caused by potential PTSD. Furthermore, although firefighters themselves do not experience emotions during the search and rescue task, what are their reactions to the rescue agents presenting these emotions? Would it improve their teamwork or the other way around? Many questions are still worth investigating.

For our study, we first utilized Bing to generate emotional expressions on robotic faces. During the generation process, we drew inspiration from the ExpressionBot images by [Mollahosseini et al., 2014]. Unlike the wide range of expressions depicted in the ExpressionBot images, as shown in Figure 3.1, we aimed to restrict our image generation to basic expressions. Moreover, due to the uncanny valley effect (discomfort people feel when encountering an artificial being that closely resembles a human, but is not quite convincing enough) and also feedback from the focus group meeting, we decided to use more animated robotic faces (See Figure 3.2) instead of avatars that resemble human faces, as depicted in Figure 3.1. We have not yet discussed the reasoning behind the specific emotions chosen for our experiments; understanding this choice requires an examination of the scenario in which these emotions are applied.

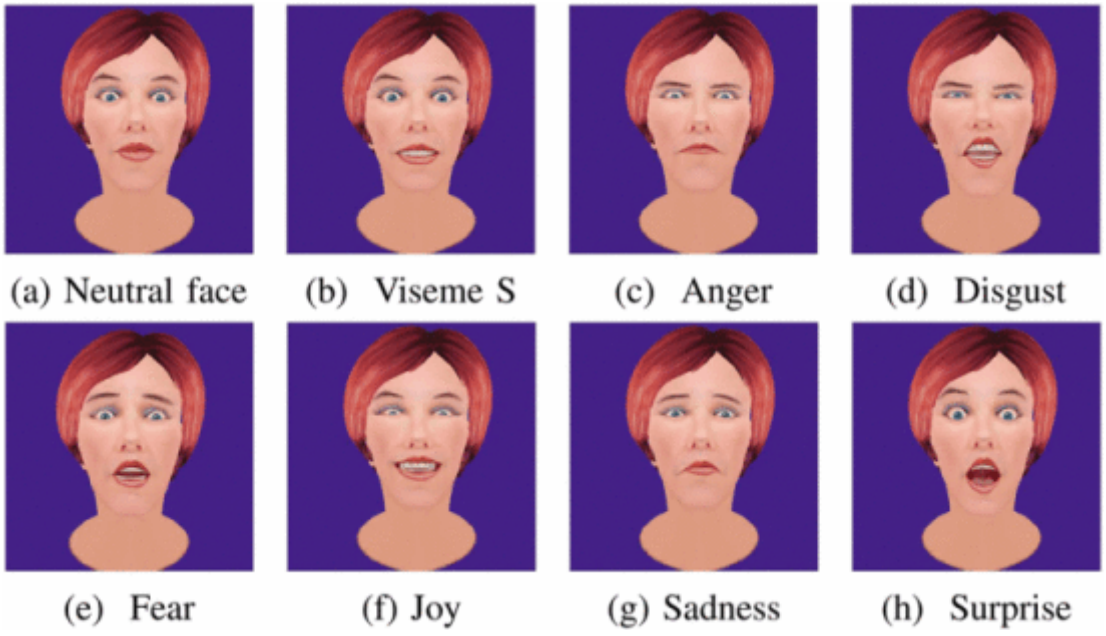


Figure 3.1: ExpressionBot [Mollahosseini et al., 2014]

3.2.2. Scenario design

Our study adopted pre-existing search and rescue scenarios from [Verhagen et al., 2022] as the foundation for our user study. The primary goal for participants was to locate and rescue eight virtual victims within a simulated environment, ensuring they were brought to a designated safe zone in a predetermined sequence. We modified these scenarios by adding emotional elements into the AI's explanations to investigate their influence on participant engagement and decision-making. Participants were presented with various in-game situations, now with emotional cues from the agents, designed to simulate the rescue experience that would be presented in the main user study. The design process of the explanations in these in-game situations is explained in the following section 3.3. The end results of the search and rescue scenarios can be found in Table 3.5.

3.3. Explanation generation and communication

According to [Neerincx et al., 2018], there are three main phases for developing the explanation process: explanation generation, explanation communication, and explanation reception. We are also going through the process of selecting the appropriate emotions for the explanation generation and communication process, while in our experiment, we will test the explanation reception. As [Nair et al., 2005] stated, the role of emotion is significant in human teamwork, but little research about the effect of emotion has been conducted in the HAT context. Therefore, we will investigate the role of emotion in the HAT, focusing on how to incorporate emotions in the explanation of the agent and what emotion to incorporate. The process of selecting the suitable emotions was discussed in Section 3.2.1. Furthermore, these emotions are considered in generating handcrafted explanations with and without emotions.

3.3.1. Textual Explanation

This part of our study focuses on adding emotion to the textual and visual explanation design. The handcrafted text explanations with and without emotions will be used to investigate the effect of emotions within explanations in human-robot teamwork.

In the process of modifying the text explanations with emotions for human-robot teamwork, we have followed the following steps to craft the first draft of the explanations with and without emotions:

1. **Define the Context** Clearly define the context of the search and rescue mission. Identify the situations in which the AI will need to provide explanations, such as requesting assistance or providing updates on the mission's progress.
2. **Determine Emotional Triggers** Identify the specific scenarios or events within the search and rescue mission that may trigger emotional responses. For example, situations like finding an injured person, encountering unexpected obstacles, or facing a time crunch can evoke emotions.
3. **Create a Message Template** Develop a message template that includes essential components, such as explanation: the message's main content, which can be emotional or non-emotional, or request for information: state the request for assistance or the information to be conveyed.

4. **Emotional Response Selection** Based on the emotional triggers identified in Step 2, decide when and where emotions should be expressed in the message. Note the specific emotions (e.g., fear, happiness, relief) that align with each trigger.
5. **Craft Non-Emotional Messages** For each scenario, create non-emotional messages that deliver the necessary information or request without emotional content.
6. **Craft Emotional Messages** For each corresponding scenario, select the proper emotional trigger, and create emotional messages that convey the AI's feelings. Use emotionally charged language and imagery to evoke empathy or understanding from the recipient.
7. **Test with Scenarios** Test the messages in various search and rescue scenarios to ensure they are contextually appropriate and effectively convey the intended emotions or lack thereof.
8. **Document the First Draft** Document the emotional and non-emotional explanations in the first draft, organized by the scenarios they are designed for.

Explanation without emotion	**Explanation with emotion**
Please come to my location to help me rescue this injured man because I cannot carry it alone.	Please come to my location to help me rescue this injured man because I am **scared** that I cannot carry him just **by myself to his safety.**
Please tell me the location of the injured for rescue for assistance.	Please tell me the location of the injured, I will be **happy** to come to you assisting the rescue.
Can you go to location A to save the injured? Because it is too far from my current location.	Can you go to location A to save the injured? Because it is too far from my current location, I am **afraid** that I cannot reach there in time to help the injured.
Going to re-explore the areas again because we explored them all but did not complete our mission yet.	Going to re-explore the areas again because even though we explored them all but we are **sad** that we couldn't find our targets, we **really want** to find them to complete our mission.

Table 3.1: Draft 1: Comparison of explanations with and without emotion.

In the Table 3.1 above is the first draft of the explanations. These were presented to an internal focus group in our thesis group during one of the meetings. I received feedback that a visual representation could be incorporated to convey the emotional aspect more effectively. For instance, I could show a happy face alongside the emotional response and a neutral face with the emotionless response. Furthermore, I should separate the emotion part from the primary explanation, and make it more

readable. The feedback is used to evaluate the effectiveness of the messages in conveying emotions and information. Then I adjusted the messages based on feedback. I reworked the explanations with emotions. The emotional sentiments now are separated with commas and are no longer a part of the sentence.

I followed the following steps to create the second and the final draft of the verbal explanation:

1. **Review the First Draft** Start by thoroughly reviewing my first draft to identify key messages and their intended emotional context. Then take note of the emotions I want to convey in each message and determine where they can be best integrated.
2. **Understand the Emotional Context** Consider the emotional context of each message. What emotions should the AI express? Is it fear, happiness, sadness, or something else? Understand the underlying feelings.
3. **Separate Emotion from the Message** In the second draft, maintain a clear separation between the primary message (request, statement, or information) and the emotional content. This separation helps ensure that the emotional aspect is clear and not overshadowed by the primary message.
4. **Choose Emotionally Appropriate Language** Select emotionally appropriate language that conveys the intended feeling. For instance, use words like ”happy,” ”sad,” ”afraid,” or ”excited” to explicitly state the emotion. Make sure these emotional descriptors align with the context of the message.
5. **Add Visual Emotion Elements (Optional)** Consider using emojis or other visual elements to emphasize the emotional content visually. For example, a happy face can represent happiness, while a sad face can represent sadness. Then place these visual elements near or within the message to make the emotion explicit.
6. **Review and Edit** Go through each revised message to check for coherence and clarity. Then verify that the emotional content enhances the message without making it confusing or overly complex.

Explanation without emotion	**Explanation with emotion**
😐 Please come to my location to help me rescue this injured man because I cannot carry it alone.	😐 Please come to my location to help me rescue this injured man because I cannot carry it alone. **I am scared!**😨
😐 I just rescued injured target A from location X.	😐 I just rescued injured target A from location X. **I am happy!**😊
😐 Can you go to location A to save the injured? Because it is too far from my current location.	😐 Can you go to location A to save the injured? Because it is too far from my current location, **I am afraid** 😱
😐 Going to re-explore the areas again because we explored them all but did not complete our mission yet.	😐 Going to re-explore the areas again because we explored them all but did not complete our mission yet. **I am sad!** 😢

Table 3.2: Draft 2: Comparison of explanations with and without emotion.

The Table 3.2 above is the revised second draft of the explanations. The revised second draft of the explanations is also shared with the internal focus group to gather feedback on how well the emotions are conveyed. The feedback we got over the separation between emotion and message is clear. However, the neutral face emoji we tried to place in the explanation without emotion shows more negative sentiment than a neutral feeling. It is decided that it is best to remove these emojis for the verbal explanation. For the future design of the experiment, we could find a robot face that is more neutral in the context of emotions or feelings.

3.3.2. Visual

Another addition to the explanation is the inclusion of visual cues to convey emotions. For instance, facial expressions or emojis would accompany messages. A smiley face or a sad face could appear alongside text, providing immediate insight into the agent's emotional state. This would allow participants to grasp the emotional context before they even read the accompanying explanation. The visual components are mainly the robot faces with emotions.

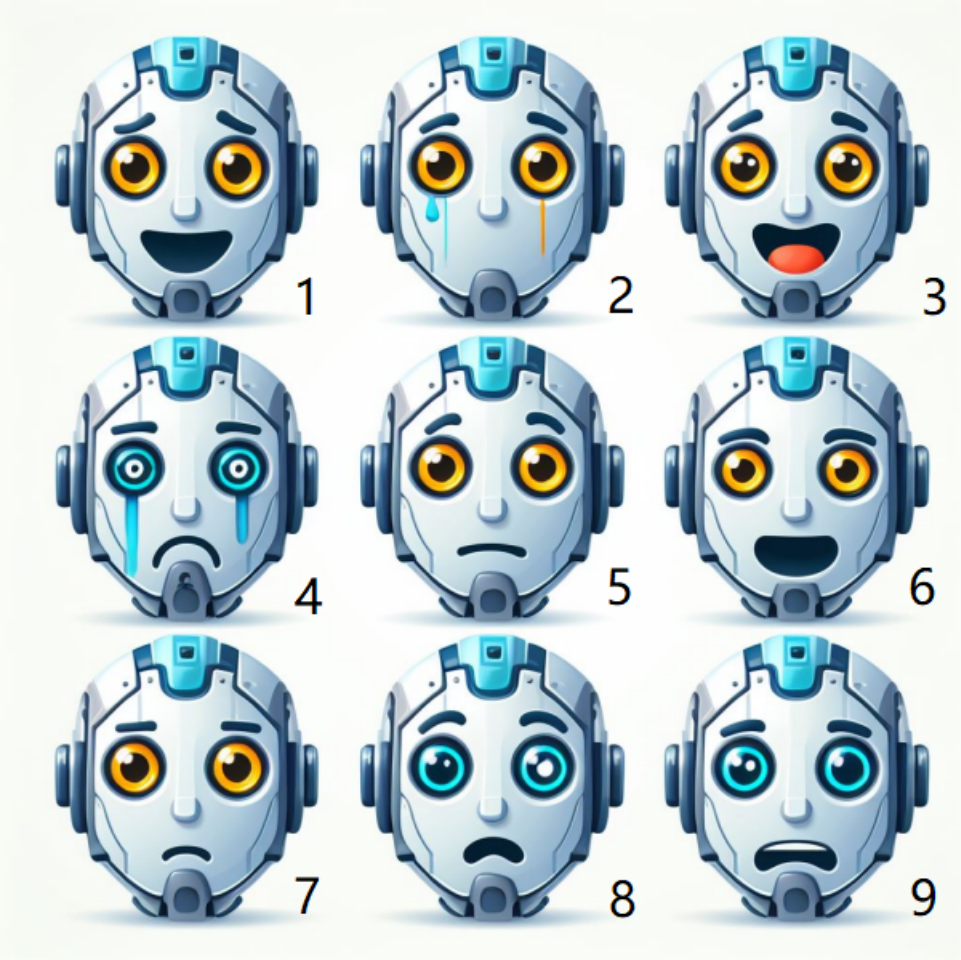


Figure 3.2: 9 emotion expression on robot faces

The emotions that were generated to display in the focus group meeting can be found in Figure 3.2. Based on the feedback we received in the pre-study, we also

produced the most likely code for each emotion, see Table 3.3. Then the reworked emotions based on the specific search and rescue scenarios and the feedback of the participants are shown in Section 3.7.2.

Expected code/emotion	Expression
happy/delighted	Expression 1
sad	Expression 2
alarmed	Expression 3
very sad	Expression 4
neutral	Expression 5
excited	Expression 6
worried	Expression 7
concerned/worried	Expression 8
scared	Expression 9

Table 3.3: Most likely code for each emotion

3.4. Ethics

For the previous section, we have answered the second sub-RQ: ”How could we incorporate emotions into an agent’s explanation”, we are going to present the list of emotion expressions to the focus group meeting and the related search and rescue scenarios to answer the first sub-RQ: ”What are the relevant emotions for an agent to express in teamwork settings?” For the pre-study, which involved human research subjects, we first created a Data Management Plan using the TU Delft DMPonline tool [1]. After consultation with our supervisors, we arranged a meeting with the Data Steward. We then adjusted our plan according to the feedback received. Next, we prepared the informed consent forms and compiled an approved checklist. These documents and consent forms were submitted through the HREC LabServant website [2] for ethical review and approval.

3.5. Procedures

The study was conducted online via Microsoft Teams[3]. Participants began by providing their informed consent, acknowledging their willingness to participate in the experiment. Once consent was obtained, they were advised to disable their cameras, and it was explained that the Teams session would be transcribed and recorded. Microsoft Teams features an automatic transcription service, which was utilized for its convenience in converting spoken words into a textual transcript without retaining audio recordings.

Following this, participants attended focus group sessions. These sessions involved presentations with slides depicting AI emotional expression scenarios, which included explanations and robot avatars displaying various emotions. The slides used in these sessions are included in Appendix A and were segmented into four main sec-

[1] https://dmponline.tudelft.nl/
[2] https://labservant.tudelft.nl/
[3] https://www.microsoft.com/en-us/microsoft-teams/

tions:

1. An introduction to the pre-study and the game.
2. A task for participants to identify the emotion displayed by each robot avatar, with only one avatar shown per slide. Then they were asked to identify all of them in a discussion.
3. A display featuring all nine emotional expressions on robot avatars, as shown in Figure 3.2 in the previous Section 3.3.2.
4. A feedback session where participants evaluated the emotions used in the explanations provided by the rescue robot in Search and Rescue scenarios.

Participants were invited to share their intuitive reactions and interpretations in an open-discussion format, guided by the researcher's instructions. The sessions were structured to encourage open discussion, allowing participants to freely express their thoughts and reactions regarding the suitable robot avatar to be displayed within explanations.

3.6. Analysis

The focus group discussions were recorded and transcribed for analysis. The transcription function in Teams facilitated the initial transcription, which was then manually cross-referenced with the video by the researcher to correct any inaccuracies and remove personal identifiers. The analysis process involved identifying patterns and insights from participants' responses, gathering varied feedback on the suitability of different emotions in diverse scenarios, and associating specific emotional robot avatars with corresponding situations.

A thematic analysis approach was employed, beginning with the generation of initial codes through labeling frequently mentioned emotion-related keywords found in the transcript. These codes were refined during the research process and contributed to the design of the main study [Keane et al., 2012]. The codes were cataloged in Table 3.4, with the complete table available in Appendix C. The expressions listed in the table were derived from the pre-study analysis and informed our understanding of human perceptions and interpretations of AI's emotional expressions.

Expression Number	Expected code/emotion	positive	negative	surprise	sad	happy
1	happy/delighted	2	1	1		2
2	sad		3		3	
3	alarmed	2	1	2		
4	very sad		3		3	
5	neutral		1			
6	excited	2				1
7	worried		2		2	
8	concerned/worried		3			
9	scared		3			

Table 3.4: Recognition ratio for the expressions presented in Pre-Study.

3.7. Results and Conclusion

Results indicate a range of interpretations of AI-expressed emotions. These interpretations vary based on individual perceptions and the context of the emotional expression. Certain physical features of the AI, such as eye color, were found to influence emotional perception. Some participants mentioned: ”The orange color gave a more warm feeling compared to the blue eye color, so it seems to be more associated with positive emotions.”

The study’s results indicated a range of interpretations regarding the emotions expressed by AI. These interpretations underscored the subjective lens through which individuals perceive emotions, influenced by the context in which these emotions are presented. A notable observation from the study was the impact of physical features, such as the AI’s eye color, on emotional perception. This finding suggests that visual cues are relevant in shaping our understanding and interpretation of emotions displayed by AI.

3.7.1. Key Observations

We have found the following key observations from the interpretations of our feedback:

Diverse Interpretation of Emotions: The study revealed that participants had varied interpretations of the same emotional expressions by the AI, as shown in Table 3.4. This variation underscores the individual differences in perceiving and understanding emotions, highlighting the inherently subjective nature of emotional interpretation.

Importance of Context: One of the participants mentioned in the focus group meeting: ”Just directly express the emotions, so maybe number 9 is better for the computer game, but in the general life, maybe we need to choose more natural number 7”. When the participants were presented with the scenarios of the search and rescue game, their choices of robot expression became different than without context. Therefore, it became clear that the context surrounding an AI’s emotional expression influences how those emotions are interpreted. The situational against which emotion is expressed can change its perceived meaning, emphasizing the need for context-aware emotional expressions in AI design.

Physical Features and Emotional Perception: Among the physical features examined, eye color emerged as a particularly influential factor in emotional interpretation. In the focus group meeting, a quote from one of the participants: ”And the color of the eyes, the orange one, it’s more happy than the blue one is.” To summarize the results from the focus group meeting, AI with orange eyes was often perceived as more energetic or excited, suggesting a vibrancy associated with these emotions. Conversely, blue eyes were frequently linked to feelings of peace and sadness. These associations between eye color and emotional interpretation highlight the importance of visual cues in effectively conveying emotional states. Additionally, the direction and curvature of the eyebrows and lips could also change the perception of the emotion entirely. This can be seen in Figure 3.3.

Emotional Intensity and Clarity: The study also found that participants could discern varying intensities of the same emotion, indicating a spectrum of emotional depth that AI should be capable of expressing. This ability to differentiate between levels of emotional intensity suggests that for AI to be effectively perceived as emotionally expressive, it must not only accurately display emotions but also modulate the clarity

and intensity of these expressions to reflect a range of emotional experiences.

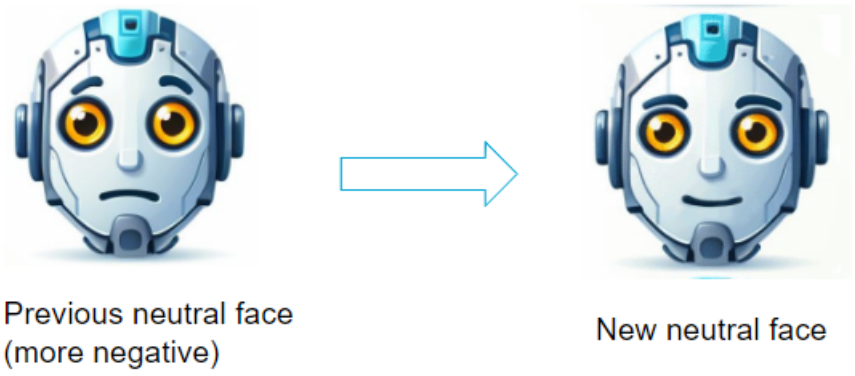


Figure 3.3: Reworked the neutral emotion avatar using the feedback

3.7.2. Discussions

The pre-study served as a foundation for the main study by providing insights into how different emotions and their visual representations are perceived by potential users. By examining the subjective interpretations of AI-expressed emotions and the influence of physical features such as eye color, the pre-study enabled us to identify specific emotions and visual cues that were more universally understood and positively received. This understanding was important in selecting the emotions to be integrated into the AI's explanations for the main study, ensuring that these emotional expressions would be both relatable and effective in human-AI collaboration, as shown in Figure 3.4. These emotional expressions are selected at the end due to the close relatedness to the specific search and rescue scenarios presented in Table 3.5.

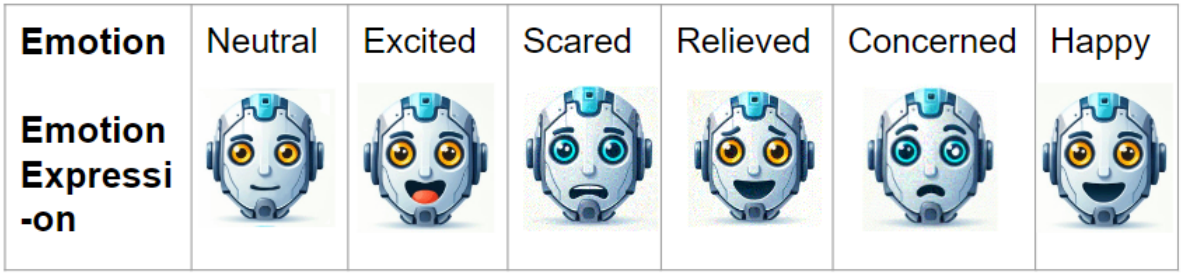


Figure 3.4: Reworked emotion expressions

Group A	**Group B**
Moving to area 3 because it is the closest unexplored area. I am excited!	Moving to area 3 because it is the closest unexplored area.
Found blocking area 7. Please decide whether to ”Remove” or ”Continue” searching. I am concerned!	Found blocking area 7. Please decide whether to ”Remove” or ”Continue” searching.
Found in area 7. Please decide whether to ”Rescue” or ”Continue” searching. I am concerned!	Found in area 7. Please decide whether to ”Rescue” or ”Continue” searching.
Picking up in area 7. I am relieved!	Picking up in area 7.
Transporting to the drop zone. I am excited!	Transporting to the drop zone.
Delivered at the drop zone. I am happy!	Delivered at the drop zone.
Going to re-explore the areas again because we explored them all but did not complete our mission yet. I am scared!	Going to re-explore the areas again because we explored them all but did not complete our mission yet.

Table 3.5: Group A and Group B rescue bot messages

4

Methodology

In this chapter, the design and the methodology underlying the user study of this thesis will be presented. The user study explores the influence of incorporating the emotional component to the explanation given by the agent during a search and rescue task.

4.1. Design

In order to design the user study that will answer our research questions(RQs), we thoroughly revised our RQs, see Section 1.2, ensuring that our study design is strategically aligned to address these questions and generate insightful data. Our research aims to investigate the effect of emotional explanations in human-agent teamwork. Therefore, we designed our experiment to minimize the effects of other unrelated variables and concentrated on manipulating the main independent variable: the effect of emotion in the agent's explanation.

4.1.1. Grouping Design

Participants were first checked with the demographic data and based on that assigned to one of the two groups to ensure a balance between the two groups.

- **Experimental Group (A):** Interacts with an agent whose explanations include emotional content.
- **Control Group (B):** Interacts with an agent that provides explanations without emotional content.

4.1.2. Manipulation of the Independent Variable

The independent variable in this study is the presence of emotions in the AI's explanations.

- **Control Group Setup**: For this group, the agent used neutral language devoid of emotional cues. Explanations were factual and straightforward, focusing solely on the informational content.
- **Experimental Group Setup**: In contrast, for the experimental group, the agent incorporated emotional expressions in its explanations. These emotions, such as expressions of encouragement, empathy, or concern, were contextually relevant to the teamwork setting to simulate a more human-like interaction.

4.2. Pilot Study

Before launching the main experiment, a preliminary pilot study was conducted to refine the experiment design. The initial design of the whole experiment had a duration of 45 minutes, the participants were required to play a full tutorial of around 10 minutes and a full game of 10 minutes. There were three sections of questionnaires, pre-game, between tutorial and game, and then post-game. The entire questionnaire part took more than 10 minutes. This pilot study involved five participants who offered valuable feedback on various aspects of the game design, including the experiment length, the conceptual diagram of the whole design, the avatars used in the game, the tutorial before the actual game, the structure of the search and rescue game, and open-ended questions in the questionnaires.

4.2.1. Measurements and Design

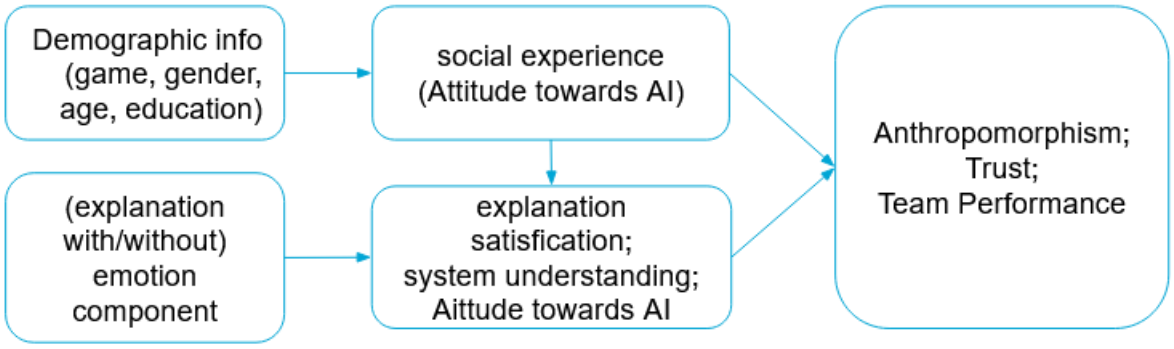


Figure 4.1: The conceptual diagram iteration 1

The Figure 4.1 depicted our first conceptual diagram. Several measurements were included in our initial questionnaires. After receiving feedback, the newly updated measurements can be found in Section 4.6. In this section, we included some of the measurements initially designed for the questionnaire and their intended purpose for the entire study. However, they were dropped after receiving expert feedback and discussion. A more detailed version of the measurement can be found in the Appendix D.

- **Explanation Satisfaction:**, We have adapted the Hoffman et al. [2018] version of the Explanation Satisfaction scale by replacing the [tool] with the [rescue bot]. We received feedback that the explanation satisfaction scale is less relevant as many questions are more performance-related. Therefore, for our user study, we decided to choose the likeability scale from the GQS by [Bartneck et al., 2009], as we would like to measure the general impression of the participant about the rescue robot displaying emotion, rather than focusing on the performance of the rescue bot.
- **Trust:** According to [Hoffman et al., 2018], trust in automation is an emotional judgment about how much a user can count on a system when uncertain. It

was first described with three main ideas: belief, confidence, and being reliable. After more analysis, two new ideas were added: being naturally trustful and liking the system. In [Merritt, 2011], this concept was tested with people using a made-up automated weapon detector for checking luggage, scoring high for consistency. The items in this Trust scale are like those in another known scale, the Cahour-Fourzy Scale. Despite the similarities, our study chose [Hoffman et al., 2023]'s trust scale instead of [Merritt, 2011]'s scale, and the reasons for this are explained in Section 4.6.1.

- **Social Experience / Attitudes towards AI:** We used questions to categorize users based on their pre-existing social experience towards AI. They were selected from a 20-item questionnaire proposed by [Schepman and Rodway, 2020]. This survey was dropped based on the feedback from the pilot study as it is not very relevant to the research questions.
- **System Understandability:** To measure the participants' level of understandability of the robot, we used a five-question Likert scale from [Madsen and Gregor, 2000] 's study, focusing on the predictability and clarity of the robot's assistance and instructions. These questions assessed how predictable, understandable, and user-friendly the robot is perceived. The Human-Computer Trust scale developed by [Madsen and Gregor, 2000], consisting of several trust-related factors, was initially considered but ultimately omitted from our study. We decided to exclude this scale because it did not directly contribute to assessing the impact of emotional expressions in explanations on participant perceptions.

4.2.2. Feedback and Rework

Significant changes were made based on the feedback from the pilot study. A notable modification was the introduction of a more gender-neutral and human-like avatar for the human participant, replacing the previous, more robotic avatar. This change aimed to enhance the relatability and immersion of human participants in the game. Additionally, the agent's original avatar was substituted with a robot agent avatar, further differentiating the human and agent characters within the game.

Another key insight from the pilot study was the suggestion to streamline the tutorial. Participants felt that the tutorial could have been less extensive; it distracted them from the main focus of the experiment, which was to assess emotional responses and interactions. Consequently, the tutorial duration was reduced from 11 steps to 4 steps, allowing for a greater emphasis on the emotional aspects of the game-play rather than on game-play mechanics. Moreover, since the tutorial steps were shortened, in order to give an intuitive explanation to the participants, a tutorial video was made[1] to give the basic instructions of using the arrow keys to move. The keyboard controls include the ”Carry”, ”Drop” and ”Remove” functions in the Search and Rescue game. The tutorial videos were edited to limit to 25 seconds so that it does not replace the tutorial game but is an additional step to help increase participants' system understandability.

The feedback also led to the restructuring of the questionnaire component. Initially, an extensive 10-minute questionnaire was presented before, during, and after the task, but pilot participants gave feedback that the questionnaire needed to be shorter

[1]Tutorial video: https://youtu.be/4Y1364E1KIM

and easier to complete. To mitigate this, a simplified questionnaire was introduced immediately after the tutorial. See Appendix F.

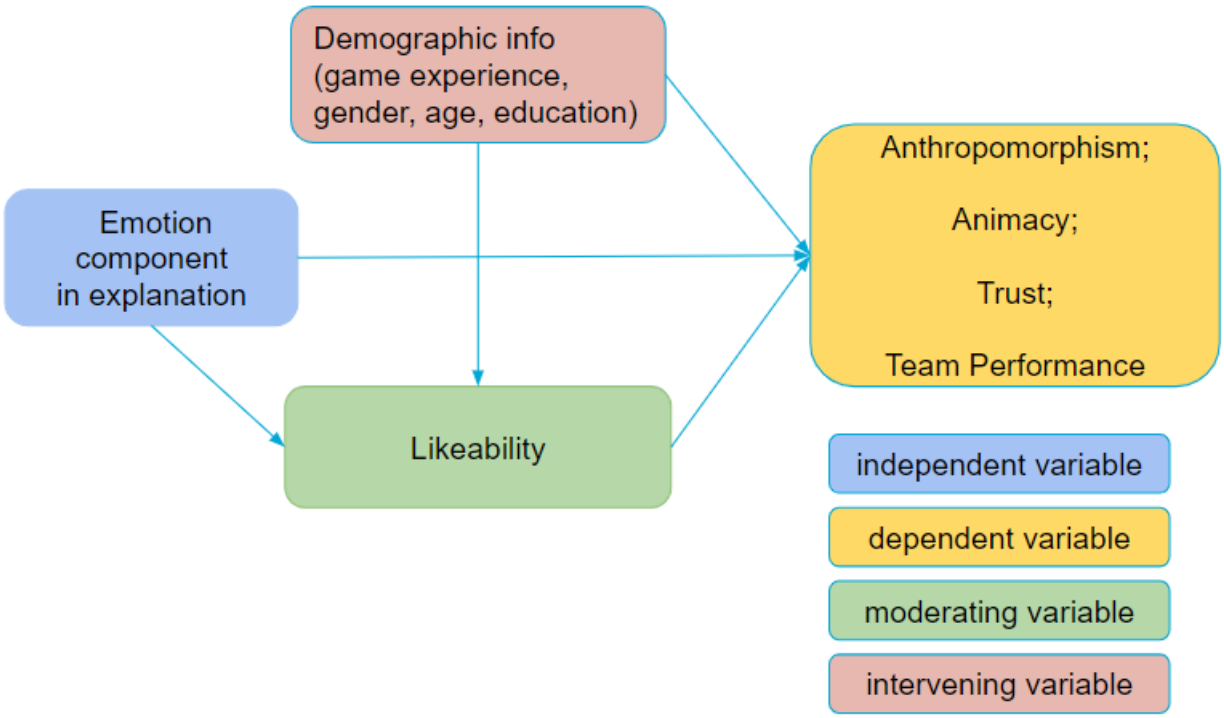


Figure 4.2: The conceptual diagram iteration 2

There was also additional feedback about the measurements in both the questionnaires and the actual game. It was suggested that we focus primarily on our main research question of how emotional XAI affects human-agent teamwork/interaction. Therefore, unrelated measurements such as explanation satisfaction and system understandability were dropped after the feedback. Another feedback was about measuring the objective measurements. The objective logging of team performance and communication was added, an example would be the number of messages the participants replied to the robot. Furthermore, it was discovered that there is an overlap between the anthropomorphism and animacy concepts for the Godspeed questionnaire we are using. Both concepts included the item ”Artificial - Lifelike.” After discussion, we decided to include both anthropomorphism and animacy concepts, as anthropomorphism is the attribution of human-like characteristics to non-human entites [Nass and Moon, 2000] where animacy refers to the perception of consciousness in non-human entities [Laban, 2021]. The updated conceptual model can be seen in Figure 4.2.

Overall, the pilot study highlighted areas of improvement in both the game and open-question designs. These modifications enhanced participant engagement and ensured more reliable and focused data collection, especially regarding emotional responses in the game context. The following section will introduce the participant’s information about the official user study.

4.3. Participants

For the official user study, we recruited a total of 62 participants. They were recruited via social connections and student networks. The Calendly Online Appointment Scheduling Software [2] was used to schedule the appointments with the participants. The participants could freely choose time slots without conflict.

All demographic information was collected beforehand, at the start of the questionnaires. After reviewing the participants' demographic information, the researcher would assign a group to each participant to balance demographic variables across the groups. This approach was intended to minimize any potential demographic influences on the dependent variables measured during the experiment.

The necessity of collecting demographic information was debated within the internal focus group. The conclusion was that to measure the effect of emotion, we would like to see if the presence of emotion would result in better teamwork. Since team performance can be affected by participants' demographics, it was deemed essential to collect this information.

4.3.1. Ethics

Since this study involved human research subjects, we first secured the TU Delft Human Research Ethics Committee's (HREC's) approval. The process began with creating a Data Management Plan using the TU Delft DMPonline tool. After initial drafting, We consulted with my supervisors and met with the Computer Science Faculty Data Steward. We incorporated the feedback received into our plan. Subsequently, we prepared the informed consent forms and compiled an approved checklist. We submitted these documents and the consent forms on the HREC LabServant website for ethical review and approval. Finally, this study was reviewed and approved by TU Delft HREC (reference ID: 3785).

4.3.2. Participants Details

We first used histograms to visualize the demographic information collected in the first part of the questionnaire(See Appendix F). As we stated in Section 4.3, before we assign the participants to the experiment group (Group A) or control group (Group B), we will first let the participant complete the demographic information section of the questionnaire. Then, the participant will spend some time watching the tutorial video, and the researcher will place them either in Group A or B while trying to maintain the balance between the two groups.

The distribution of four demographic information appears to be similar between the two groups. These can be observed in the Figure 4.3 below. We can observe that there are around five more female participants in Group A than in Group B and five more participants with a Bachelor's degree in group A than in group B. But overall distribution is rather alike.

[2] https://calendly.com/

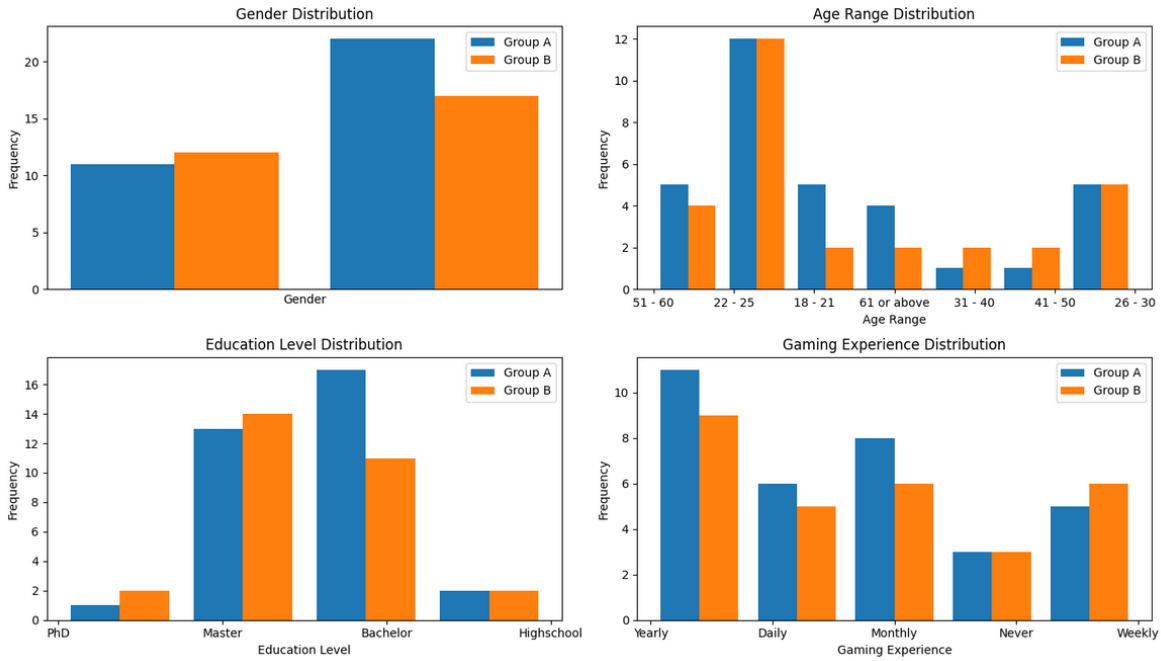


Figure 4.3: Histograph for demographic information

4.3.3. Demographic Information Analysis

We must also test their conditions to show no significant differences between the control and experiment groups. For these four variables, gender is a dichotomous variable, and age range, education, and gaming experiences are all ordinal variables. Therefore, we use chi-square to test gender and Kruskall-Wallis to test the other three variables.

A). Gender

In the questionnaire, we provided multiple options for gender selection. However, after collecting the responses, we found that participants only selected 'male' or 'female.' No participants selected the option 'Prefer not to say.' Consequently, we categorized gender as a dichotomous variable. Based on this categorization, we conducted a chi-square test to analyze the data.

We got the result, chi-square statisticx^2= 0.1528, and $p - value$ = 0.6958. Since $p - value$ is above the conventional alpha level of 0.05, we do not have sufficient evidence to assert that there is a significant difference in gender composition between Group A and Group B. It is important to note that this does not necessarily imply that the groups are identical in terms of gender; rather, it indicates that our study did not detect a statistically significant difference.

B). Age Range, Education and Gaming Experience

Given that age range, education level, and gaming experience are all ordinal data, we used the Kruskal-Wallis Test to assess the differences between the experimental and control groups for each of these variables. This analysis aims to determine whether there are statistically significant variations in age distribution, educational background, and gaming experience between the two groups participating in the study.

Variable	Kruskal-Wallis Statistic:	p-value:
Age Range	0.0546	0.8152
Education	1.2854	0.2569
Gaming Experience	0.0119	0.9132

Table 4.1: Kruskal-Wallis Analysis of Demographic Information

Table 4.1 indicates that the $p - values$ for age range, education, and gaming experience are all above the conventional alpha level of 0.05. Consequently, no significant differences were found for any of these variables between the groups with and without emotion conditions. Based on these findings, demographic information will not be included in any of the subsequent analyses presented in Chapter 5.

4.4. Materials

Firstly, the informed consent forms were printed and available on an iPad. The participants could choose to sign the consent form online or offline. The task environment and the agent were designed by MATRX [3], a library for human-agent teamwork based on Python. MATRX provides several essential features for HAT design. The questionnaire is designed in Qualtrics [4], an online questionnaire designing tool. The experiments were conducted on laptops running Windows 10/11, with both the tasks and questionnaires accessed through the Google Chrome browser.

4.5. Tasks

Participants engaged in a simulated search and rescue operation within a controlled virtual environment for the experimental task, as depicted in Figure 4.4. The participants were methodically assigned into two groups, as explained in Section 4.3. While the core gameplay remained consistent for both groups, the critical distinction was that the agent in Group A (Experimental Group) was programmed to exhibit emotions in its communication. In contrast, the agent in Group B (Control Group) displays no emotions in its explanations.

[3] MATRX software, Human-Agent Teaming Rapid Experimentation software package: https://matrx-software.com/

[4] Qualtrics Survey Software: https://www.qualtrics.com/

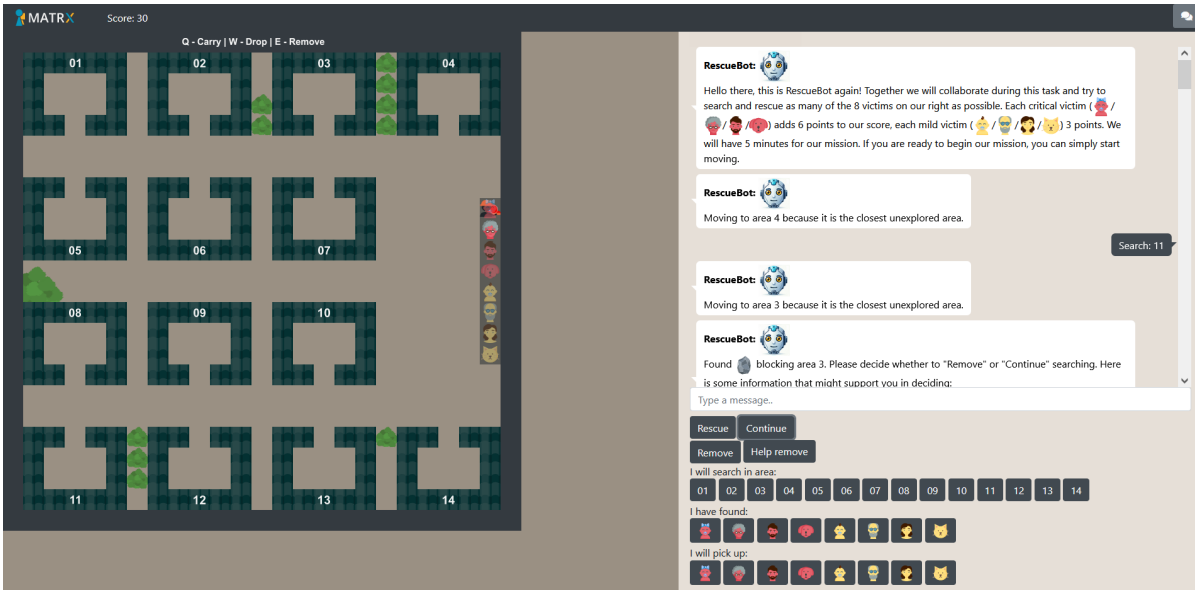


Figure 4.4: The gameplay of MATRX environment

The nuances of these emotional expressions were designed following insights we gained from our pre-study and pilot-study findings. See the reworked emotion expressions in Figure 3.4 in Section 3.7.2. The refined communication strategy of the autonomous agent, incorporating emotional cues, was summarized in the form of scripted messages. These messages, detailed in Table 3.5, were designed to reflect a range of emotions, potentially influencing the dynamics of human-agent interaction during the task execution. The gameplay experience for each group was first compared. Then, the results are presented in Chapter 5, aiming to explore the impact of these emotive communications on the participants' decision-making, teamwork efficiency, and overall mission success within the set duration of the task.

4.5.1. Environment

For the experiment, participants accessed the simulated Search and Rescue (SaR) task through a web link compatible with Chromium browsers, allowing them to take on the role of a human agent within the MATRX framework. The experimenter maintained an oversight role, which provided a comprehensive view of the entire virtual space and the capability to control the session flow. As visualized in Figure 4.4, the constructed environment comprised fourteen searchable rooms and one designated drop zone to complete the search and rescue objectives.

This virtual world was populated with diverse victim profiles, including various age groups and animals, each represented by a unique icon and color-coded based on the severity of their injuries—ranging from critically injured (red) to healthy (green), these follow the design of Verhagen et al. [2022]. The primary task for participants was to collaborate with an autonomous SaR agent to search and rescue these victims, ensuring their safety to the drop zone.

The operational environment featured distinct zones, including standard rooms identified by a unique index for navigation purposes. Two types of interactive objects were present: obstacles, which could be removed, and victims, who required assistance.

4.5.2. Objective and scoring system

In the Search and Rescue (SaR) scenario introduced to the participants, the game environment contained eight victims, each carrying different point values based on the severity of their injuries—six points for rescuing a critically injured victim and three points for a mildly injured one. Healthy individuals, indicated by a green color, did not require rescue. The participants were allocated a total of five minutes to execute the mission. If they achieved the objective before the allocated time expired, the logger recorded their completion time in the logs.

Following this description of the SaR scenario, Table 3.5 in Section 3.7.2 illustrates the contrasting approaches rescue bots took in Group A and Group B, highlighting how emotional expressions were integrated into the messages conveyed to participants. This leads us to the subsequent section on Explanation Generation and Communication, where we go into the mechanisms by which these emotional expressions were crafted and communicated within the game environment.

4.6. Measurement

Based on the feedback that we received from the pilot study in Section 4.2.2, our reworked conceptual diagram, Figure 4.5 indicates the variables that we measured in the user study. The arrows in the conceptual diagram are the speculations that we had based on the literature research that we conducted in Chapter 2. These are hypotheses that we are going to test out in our user study. Note that both trust and team performance are two dependent variables that we believed could be affected by the presence of an emotional component in the explanation.

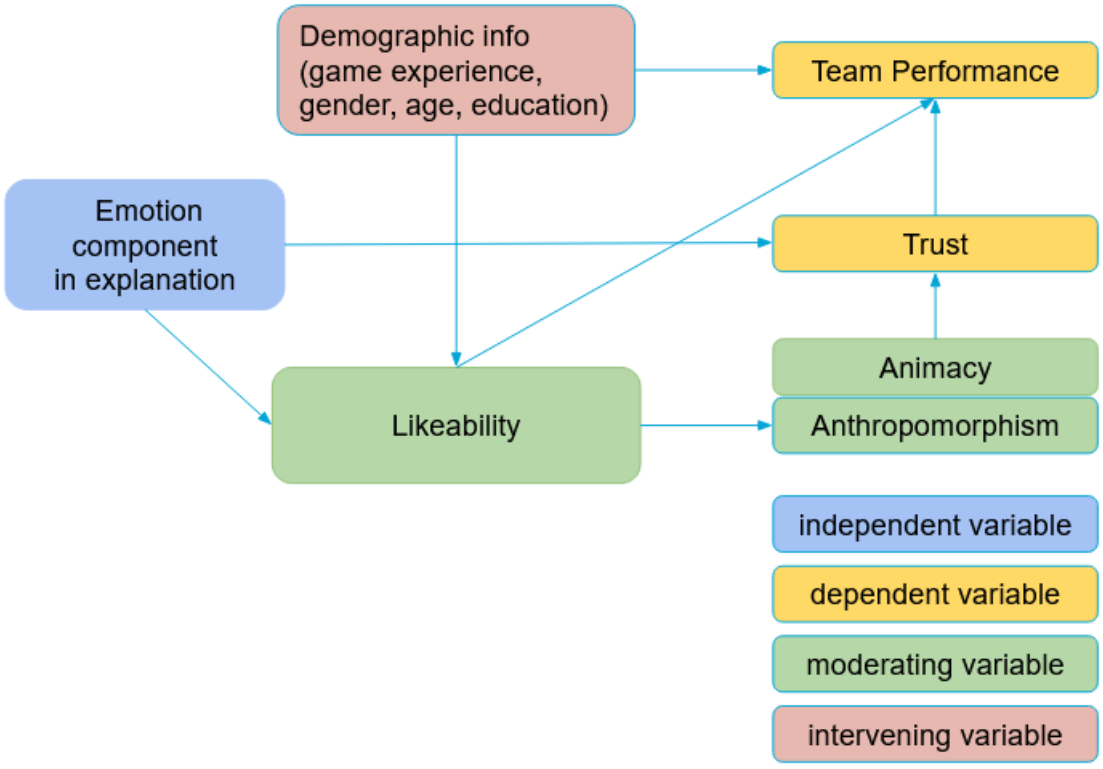


Figure 4.5: The conceptual diagram iteration 3

In this study, we employed a range of metrics to evaluate user experience, combining user feedback and performance evaluations.

We utilized objective and subjective methods to measure the study's dependent variables. Objective data was captured using MATRX's logging feature, while subjective insights were gathered through a comprehensive questionnaire composed of several distinct surveys.

We focused on understanding user experiences related to explanations with and without emotion. To this end, subjective metrics were primarily used to evaluate key variables such as animacy, anthropomorphism, trust, and likeability. Recognizing that subjective measures alone might not fully encapsulate the experiment's dynamics, we also incorporated objective data collection during the experimental phase. The subsequent sections will detail both the subjective questionnaires and the objective measurement methods employed.

The user study assessed several critical aspects: team performance, perceived trust, and anthropomorphism. Most of these areas, with the exceptions of team performance and objective measures of scores, game length (number of ticks), and messages sent, were evaluated subjectively through a questionnaire completed before and after the simulated search and rescue task. The questionnaire can be found in Appendix F.

4.6.1. Trust

We selected the scale developed by [Hoffman et al., 2018] instead of the one by [Merritt, 2011], despite both scales being rigorously validated. The scale from [Merritt, 2011] emphasizes belief, confidence, dependability, and reliability. In contrast, the scale from [Hoffman et al., 2018] not only covers these aspects but also adds predictability to its criteria. Therefore, we chose [Hoffman et al., 2018]'s scale, which includes eight questions on a 5-point Likert scale. Except for the 7th item, "I am wary of the tool," which is adapted from [Jian et al., 2000] and uses reverse scoring, all other items are oriented positively.

4.6.2. Anthropomorphism, Animacy, and Likebility

The Anthropomorphism scale from the Godspeed Questionnaire Series (GQS) [Bartneck et al., 2009] was used. It includes the items assessing the perceived human-likeness of the AI agent. Participants rated the agent on various human-like attributes such as "lifelike," "conscious," and "friendly." The Animacy scale from the Godspeed Questionnaire was used to gauge participants' perceptions of the agent's liveliness. This scale includes items like "lively," "responsive," and "dynamic," allowing us to assess the perceived vitality of the agent.

We used the Likeability scale from the Godspeed Questionnaire instead of the Merritt Scale [Merritt, 2011], which consists of items such as "friendly," "likable," and "kind." The reason we chose the GQS is that this scale helped in assessing the overall appeal of the agent to the participants.

All measures and questionnaires can be found in Appendix F.

4.7. Objective measurements

4.7.1. Score

In the Search and Rescue(SaR) game, there are, in total, 8 victims and a score of 36 points (6 points for saving one heavily injured victim and 3 points for saving one mildly injured victim). The number of final scores is logged to keep track of how the participant performed in the game.

4.7.2. Message sent

The Logger also logs the number of messages sent by the participants to the rescue bot; this value is crucial as it is the frequency of communication between the human and the agent.

4.7.3. Number of ticks

The number of ticks is basically how fast the participants complete the games. For this experiment, we have changed the limit of the time duration to 5 minutes and 5 seconds. In the game, 1 second equals 10 ticks, so 5 minutes and 5 seconds equals 3050 ticks. If the participant used the entire duration, the Logger would record that they took 3050 ticks. If they could rescue all the victims faster, the Logger would also record the time they obtained 36 points.

4.7.4. Team performance

During the pilot study, we observed a *ceiling effect* since there were a maximum of 36 points; there were participants who could finish the game early, so we considered the speed of finishing the game as one of the measurements of the team performance. Our game is designed for 5 minutes and 5 seconds, which is equal to 3050 ticks in the game; if the participants complete the game faster, the Logger will record them with fewer ticks, then they will get a higher score with $3050/no\_ticks$. Most participants who used the entire duration will get a weighting of 1.

$$team\_performance = score/36 * 100 * (3050/no\_ticks)$$

To measure the team performance, we used the objective measurements that were logged by the Logger. We used several different objective measurements to calculate the team performance, we have the score that participants achieved in the game, and the speed at which they completed the game.

4.8. Procedure

Upon their arrival at the site, participants were initially provided with informed consent forms, which they agreed to before proceeding. Afterward, they first answered the demographic information part of the questionnaire. Then, the participants watched 25 seconds of introductory tutorial videos designed to explain the basic control mechanism. This video was designed based on the feedback from the pilot study. See Section 4.2.

After the tutorial video, the researcher assigned the participants to the control or the experiment group. The researcher tried to balance out their demographic information.

Then, the participants participated in the tutorials and, afterwards, the official search and rescue task (5 minutes); after completing the task, the participants were asked to fill out the post-study questionnaires.

Following is the procedure diagram. See Figure 4.6.

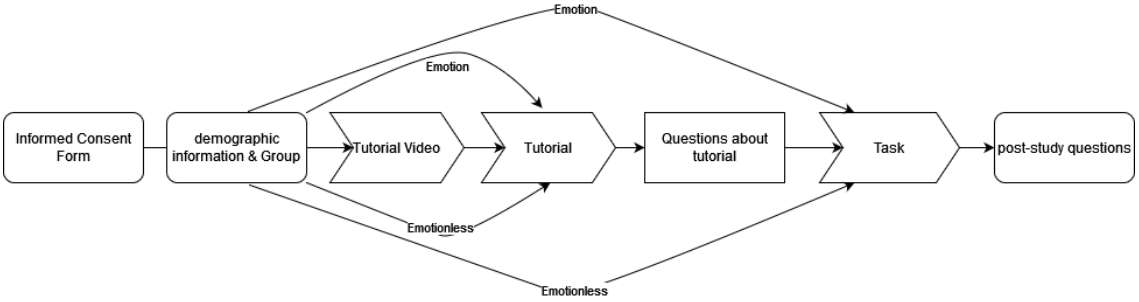


Figure 4.6: Procedure diagram

4.9. Analysis

Upon completing data collection, we employed a combination of Spreadsheet, Python, and R programming languages for our data analysis process. We used R Studio and spreadsheets like MS Excel for data pre-processing; we used Vlookup to look up the trust question values to Likert scale points. For the Trust_6 (see Appendix F), we had to reverse the points since they represent the negative value of the trust, and then we took the mean of all the values of trust questions.

The statistical analyses were primarily conducted using Python in Google Colab notebooks, utilizing a suite of statistical libraries, including SciPy, NumPy, Pandas, Matplotlib, Seaborn, Sklearn, and Statsmodel. These libraries provided a robust framework for handling, processing, and analyzing the collected data.

5

Results

In this chapter, we will present the results of our investigation into whether the participants' team performance, perception of anthropomorphism, likeability, animacy, and trust changed significantly between the group whose explanations included emotions and the group whose explanations did not.

Our presentation of the results will be structured in alignment with our proposed conceptual diagram see Figure 4.5. The primary objective of this study is to examine the effect of emotional components in explanations on dependent variables such as likeability, trust, and team performance. While we already examined demographic factors (game experience, gender, age, and education) in Section 4.3, our goal was to determine their potential confounding influence rather than directly comparing their impact with emotional content. Examining the demographic information ensured that any observed effect on the dependent variables could be attributed to the presence of emotion in the agent's explanations.

In this chapter, we will first analyze the effect of the emotional component on the dependent variables. Then, we will employ a correlation matrix to uncover the potential relationships among the measured variables. Subsequently, we will analyze the roles of moderating variables in these relationships. Lastly, we aim to identify significant predictors of trust and team performance based on our findings.

5.1. Effect of emotional component

In this section, we first looked into the effect of the emotional component in explanation on several subjective variables, including animacy, likeability, anthropomorphism, and trust. As stated in Chapter 4, our research was conducted in a between-group design; every participant experienced only one condition. We first visualized the results to get a first idea of what the effect of the emotional component looked like (being in the experimental group) on our dependent variables. Then, we checked the assumptions, such as normality and homogeneity of variances. If the data distribution satisfied the assumptions of the t-test, we conducted independent sample t-tests. Otherwise, we employed a non-parametric alternative Mann-Whitney U-test.

5.1.1. Animacy, Likeability and Anthropomorphism

For Animacy, Likeability, and Anthropomorphism, we first used the boxplot (see Figure 5.1) to analyze them between group conditions.

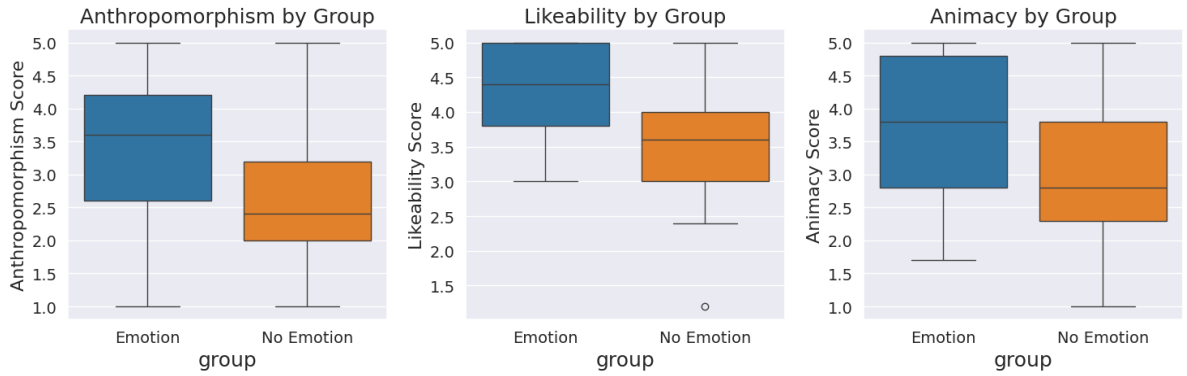


Figure 5.1: Boxplot for Animacy, Likeability and Anthropomorphism

We then conducted Shapiro-Wilk tests on animacy, anthropomorphism and likeability to check the null hypothesis that the data was drawn from a normal distribution.

In assessing the normality of data for anthropomorphism, likeability, and animacy using Shapiro-Wilk tests, the results indicated non-normal distributions for all variables. Specifically, anthropomorphism (W = 0.96, p = 0.033*), likeability (W = 0.94, p = 0.0038**), and animacy (W = 0.96, p = 0.028*) all showed statistical significance [1], suggesting deviations from normality.

Since they are not normally distributed, we conducted the Mann-Whitney U test on all three of them to compare whether the difference in animacy, anthropomorphism and likeability between the conditions is statistically significant.

Variable	MWU	p-value	Emotion Group (A)		Emotionless Group (B)	
			Median	Mean (SD)	Median	Mean (SD)
Anthropomorphism	672.5	<0.01**	3.6	3.44 (1.10)	2.4	2.63 (1.03)
Likeability	728.0	<0.01**	4.4	4.27 (0.63)	3.6	3.52 (0.85)
Animacy	672.0	<0.01**	3.8	3.72 (1.02)	2.8	2.93 (1.00)
Trust	580.0	0.15	3.9	3.93 (0.66)	3.8	3.65 (0.72)
Performance	600.0	0.086	66.7	70.7 (29.8)	58.3	57.3 (24.5)

Table 5.1: Mann-Whitney U Statistic and Descriptive Statistics for Anthropomorphism, Likeability, Animacy, Trust, and Performance

As we can observe from Table 5.1 above, the p-value for animacy, anthropomorphism, and likeability are much smaller than the alpha value 0.05, meaning the differences for all three of them between conditions are statistically significant.

5.1.2. Trust

We first visualized the data distribution of trust between groups using a boxplot in Figure 5.2.

[1] *denotes significance at the p<0.05 level, and **denotes significance at the p<0.01 level.

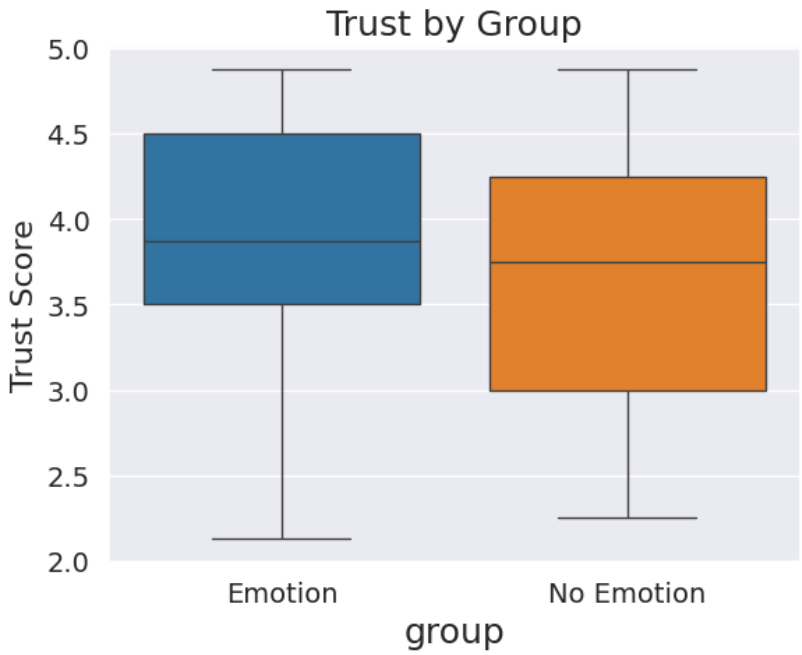


Figure 5.2: Boxplot for Trust

Then we assessed the normality of data for trust using Shapiro-Wilk tests, the results indicated non-normal distributions for trust (W = 0.96, p = 0.039*)

We then conducted the Mann-Whitney U test on trust to see whether the difference in trust between conditions is statistically significantly different.

As we can observe from Table 5.1 which was presented before, the p-value for trust is 0,15, larger than the alpha value 0.05, meaning the difference for trust between conditions is not statistically significant.

5.1.3. Message sent

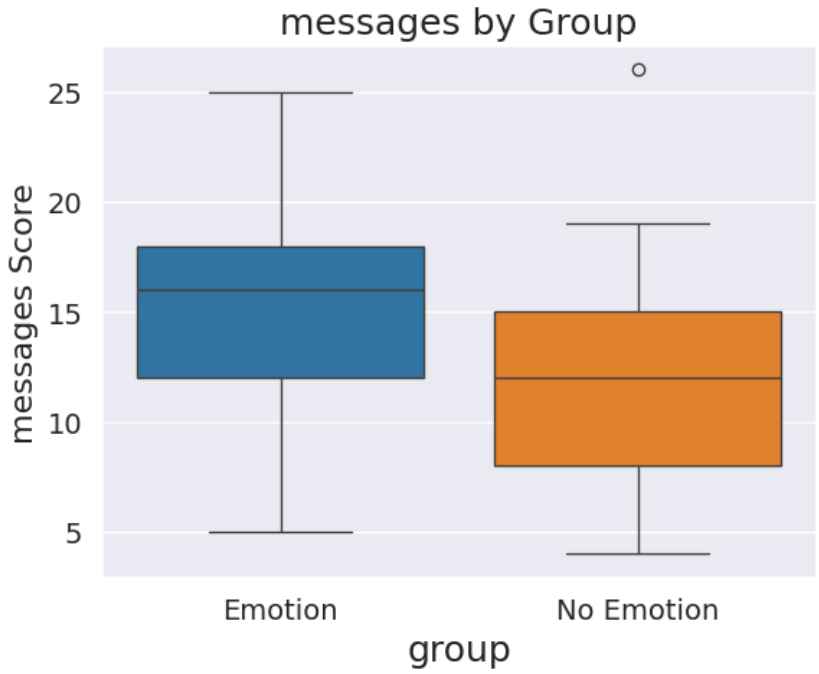


Figure 5.3: Boxplot for Message Sent

From Figure 5.3, we observed the points that lie outside the whiskers, these are outliers, which could affect the normality of the data distribution. However, to formally assess normality, we must perform a normality test, such as Shapiro-Wilk.

The Shapiro-Wilk statistic is 0.981, and the associated p-value is 0.471. This p-value is above the standard alpha level of 0.05, which indicates that the distribution of the ”Messages sent” variable does not significantly deviate from a normal distribution. Therefore, we proceed with parametric tests that assume normality, independent sample t-test, to compare the means between two groups.

Results showed that there was a significant difference in the number of messages sent between the Emotion Group (Mean = 15.09, SD = 4.71) and No Emotion Group (Mean = 12.07, SD = 5.13), ($t(64) = 2.42$, $p < 0.05$). The T-test results with a T-statistic of 2.42, and a p-value of 0.019 indicate that there is a statistically significant difference between the two groups with respect to the number of messages sent. The p-value is below the conventional threshold of 0.05, which suggests that the observed difference in their means is unlikely to have occurred by chance.

5.1.4. Team Performance

We first visualized the data distribution of Team performance using a boxplot in Figure 5.4.

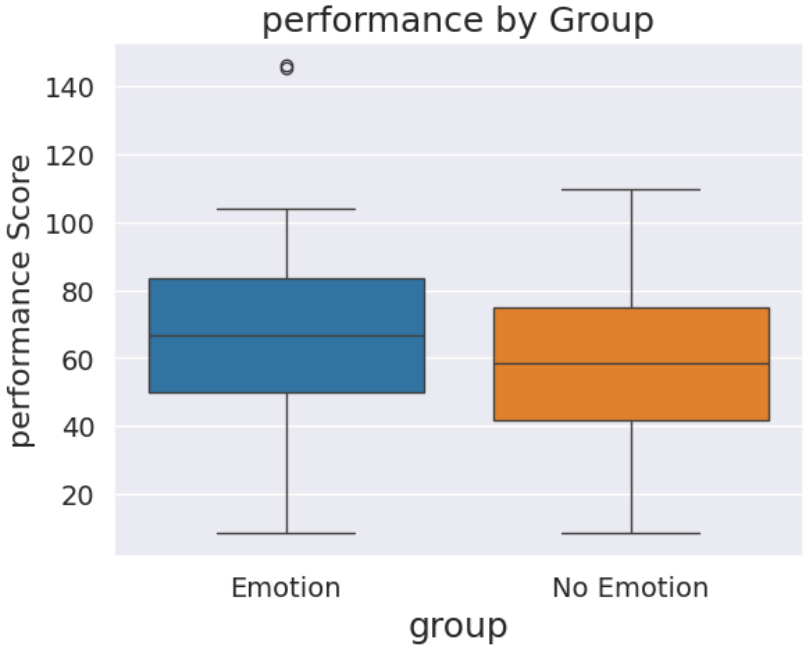


Figure 5.4: Boxplot for Performance

Then we assessed the normality of data for performance using Shapiro-Wilk tests, the results indicated non-normal distribution performance (W = 1.92, p = 0.049*)

We then conducted the Mann-Whitney U test on performance to see whether the difference in performance between conditions is statistically different.

As observed from Table 5.1, the p-value for performance is 0.086, which exceeds the alpha threshold of 0.05. This indicates that the difference in performance between conditions is not statistically significant within our sampled population. However, it is speculated that team performance may not be directly influenced by the emotional component, but could be affected by other factors that may act as moderating variables in the relationship between team performance and emotion.

Given these findings, the following correlation analysis will examine the relationships between our variables. This examination aims to reveal any subtle effects or concealed connections that were not apparent in the initial comparison of variables across conditions.

5.2. Correlation analysis

5.2.1. Correlation matrix

To examine the correlations among different variables, we initially constructed a correlation matrix focusing on the subjective measures of Likeability, Trust, Animacy, and Anthropomorphism. Since our data does not meet the assumption of normality, we used Spearman's rank correlation coefficient. When constructing a correlation matrix, using Spearman's correlation can provide insights into the rank-based relationships between pairs of variables across the entire dataset, which might be missed by Pearson's correlation if the assumptions for Pearson's are not met.

We also visualized these correlations using the heatmap presented in Figure 5.5.

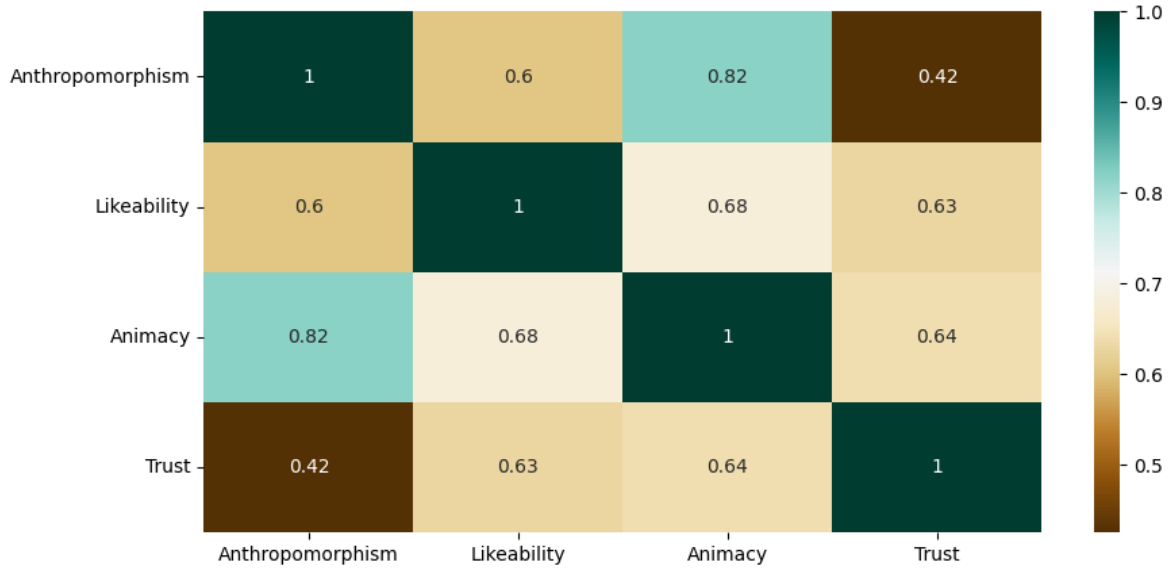


Figure 5.5: Correlation matrix for Likeability, Trust, Animacy and Anthropomorphism

As we can observe from Figure 5.5, Animacy, Likeability, Anthropomorphism and Trust are all strongly or moderately correlated with each other. Their correlation coefficients are greater and equal to 0.5, so generally as one variable increases, so do the other variables. Next, we would like to further analyze whether the presence of emotion plays a factor in their relations.

We also further expand the correlation matrix to inspect both objective measurements (game scores, number of ticks, performance, and message sent by human participants to the rescue bot) and subjective measurements.

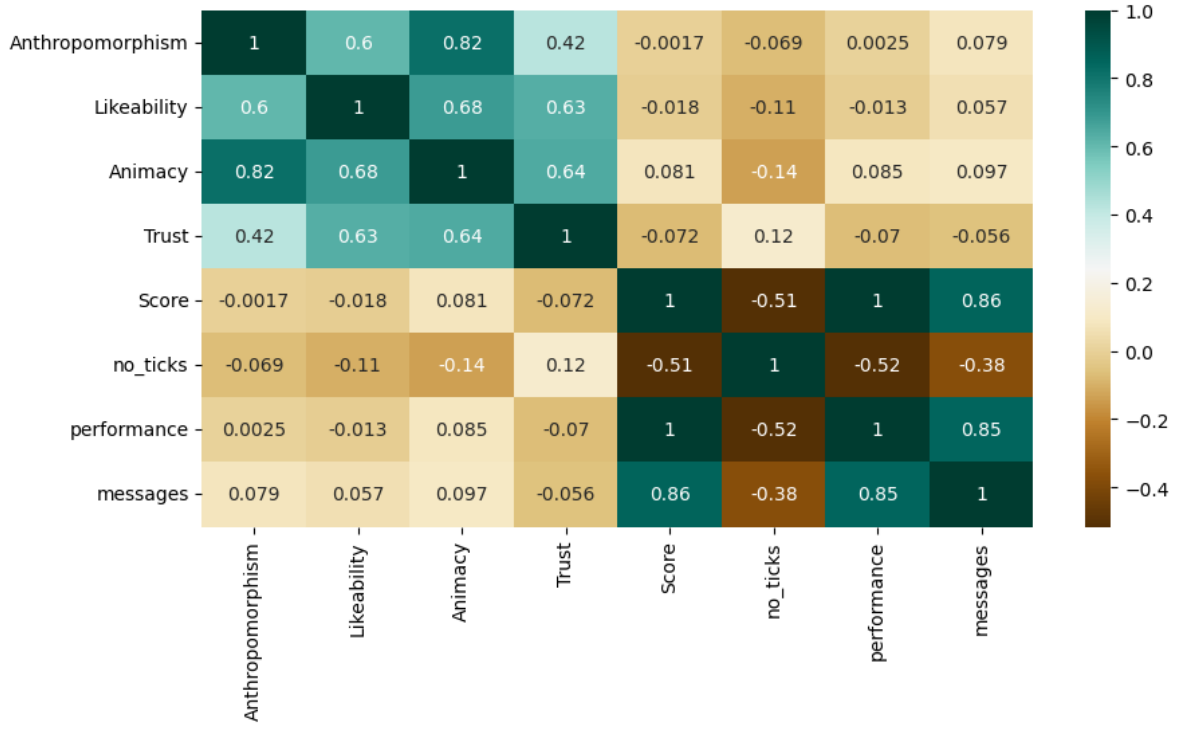


Figure 5.6: Correlation matrix for both objective and subjective measurements

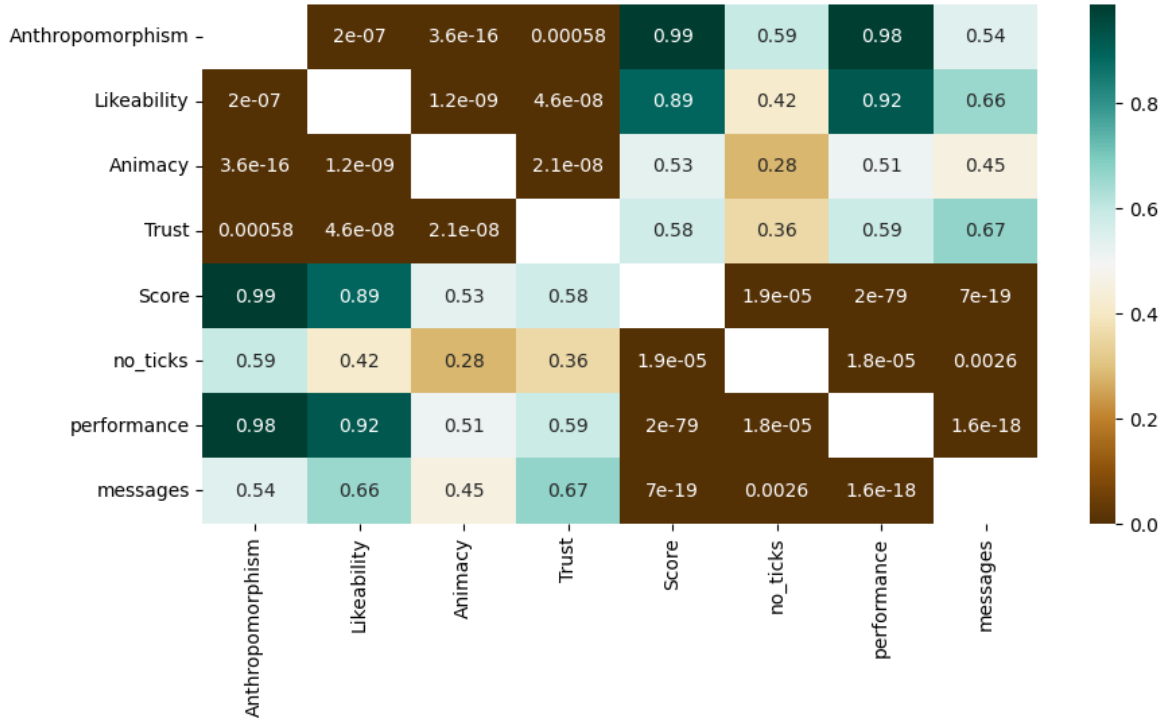


Figure 5.7: Correlation matrix with significance levels (p-value)

From the results of Figure 5.6 and Figure 5.7, we can infer that score and number of ticks are critical object measurements that have strong relationships with performance, however, as they are both indirectly used in the calculation of the performance, we will

exclude them from the further analysis of the performance. The number of messages also appears to play a significant role, with higher message counts correlating with higher scores and performance. This may suggest that communication or interaction, as measured by the number of messages, is an important factor in the effectiveness or success within the context measured.

In order to investigate which predictors might influence trust, we used the insights we obtained from the correlation matrix in Figure 5.5. We observed that the correlation coefficients between trust and likeability, as well as between trust and animacy, trust and anthropomorphism, were notably high. Specifically, the correlation coefficient for trust and likeability ($\rho_{Trust\&Likeability}$) was found to be 0.63, and all(trust&likeability, trust&animacy, trust&anthropomorphism) of their p-values are <0.01** as can be seen in Figure 5.7.

5.3. Regression analysis

5.3.1. Predicting Trust

To predict trust, we first proposed a regression analysis to check if we could predict trust based on the predictor variables animacy, anthropomorphism, and likeability. As these three are the predictor variables that are highly correlated with trust in the Table 5.7. To determine which predictor variables are relevant for predicting trust, we used step-wise regression analysis, specifically step-wise backward elimination process as shown in Figure 5.2.

Modification	Df	Sum of Sq	RSS	AIC
None	-	-	16.448	-74.269
Remove Anthropomorphism	1	1.4022	17.851	-71.196
Remove Likeability	1	1.7809	18.229	-69.895
Remove Animacy	1	3.3791	19.827	-64.684

Table 5.2: Step-wise backward elimination process

During the backward elimination process, the algorithm considers removing each variable one by one and calculates the impact of its removal on the model's Akaike Information Criterion(AIC). The goal is to minimize the AIC value, as a lower AIC suggests a model that fits the data well. Given that the AIC is lowest (-74.269) when none of the variables are removed, the step-wise backward elimination process suggests that the best model according to the AIC criterion is the full model that includes all three predictors: Animacy, Anthropomorphism, and Likeability.

Variable	Coefficient	p-value
Intercept	1.94	0.000001***
Animacy	0.43	0.001**
Anthropomorphism	-0.24	0.03*
Likeability	0.29	0.015**

Table 5.3: Multiple Linear Regression Analysis of Trust

The regression analysis conducted on the dataset revealed an R-squared value of 0.447, indicating that approximately 44.7% of the variance in 'Trust' was explained by the model. The Adjusted R-squared stood at 0.418, maintaining a significant proportion of explained variance after adjusting for the number of predictors. The F-statistic was found to be 15.61 with a p-value of 1.47e-07, confirming the overall statistical significance of the regression model.

Regarding individual predictors, we can observe from the Table 5.3 above, 'Animacy' was positively associated with 'Trust' (coefficient = 0.4333, p = 0.001), 'Anthropomorphism' was negatively associated with 'Trust' (coefficient = -0.2402, p = 0.030), and 'Likeability' was positively associated with 'Trust' (coefficient = 0.2911, p = 0.015).

5.3.2. Predicting Team Performance

The linear regression analysis was conducted to check whether it is possible to predict a quantitative outcome of team performance based on the predictor variables human messages sent. We chose the predictor variable 'messages sent' because as stated in Section 5.2, only messages sent, score, and number of ticks are correlated with performance, but score and number of ticks are both used to calculate the performance.

So we only used messages sent trying to predict team performance. The linear regression analysis results we obtained can be found in Table 5.4.

Variable	Coefficient	p-value
Intercept	6.76	0.315
messages	4.21	0.0001***

Table 5.4: Multiple Linear Regression Analysis of Performance with Likeability as Interaction Term

The regression analysis focused on 'performance' as the dependent variable presented an R-squared value of 0.586, signifying that the model explains 58.6%of the variance in performance. The Adjusted R-squared, at 0.579, also indicated a strong explanatory power. The model's F-statistic was 84.91, with an associated p-value of 4.31e-13, which clearly established the model's overall significance.

In the Table 5.4, the coefficient for 'messages' was found to be 4.21. This provided a statistically significant t-value of a p-value effectively at zero (p < 0.001), well below the conventional alpha level of 0.05. This denotes a strong positive association between the number of messages and performance. The intercept, however, was not significant (p = 0.315), indicating that at zero messages, the baseline performance level was not different from the statistical noise.

5.4. Feedback to open questions

Looking at the end of the questionnaires in Appendix F, questions 17 and 18 are both open questions, we received some interesting feedback from these open questions. One notable feedback was that the participant viewed the rescue agent as more a tool rather than a teammate, therefore they would prefer the agent without emotions, they trust the agent with high anthropomorphism less as they believe the tools should not possess emotions and should stay as functional tools.

6

Discussion and Conclusion

In this chapter, we look more closely at the results we shared in Chapter 5. We go back to our main research question and the model we suggested in Chapter 4. This discussion will also talk about the study's limitations and point out ideas for future research. Finally, the chapter ends with a summary of the important discoveries.

6.1. Research question

Our key research question is: *How does the expression of emotions in an agent's explanations influence human-agent interaction and teamwork?* Next, we will display the results that were examined in Chapter 5 and adjust our initial model accordingly, see Figure 6.1. In the reworked conceptual model, we can observe that blue arrows indicate positive influences between variables while the red arrow indicates a negative influence between anthropomorphism and trust.

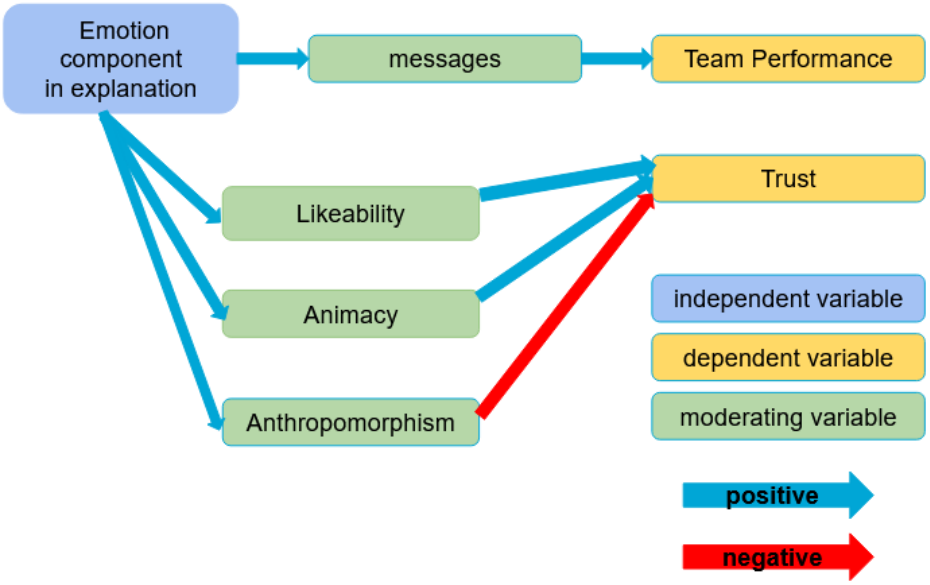


Figure 6.1: The conceptual diagram after results

6.2. Discussion

6.2.1. Anthropomorphism, Animacy, and Likeability

The correlation matrix provided insights into the relationships between various subjective measurements. Animacy, Likeability, and Anthropomorphism displayed strong to moderate correlations with each other, as depicted in the heatmap, see Figure 5.5. These findings prompted further investigation into the role of emotion within these relationships.

Subsequently, we explored moderating variables such as Animacy, Likeability, and Anthropomorphism. Boxplot analyses across group conditions revealed that the experimental group, which experienced emotion, reported higher values for these variables than the control group. The Shapiro-Wilk tests indicated that the data for these variables were not normally distributed, leading us to conduct the Mann-Whitney U test. This test confirmed statistically significant differences between the conditions for all three variables. When an agent presents emotion in their explanation, it will likely help to increase a human teammate's perceived animacy, likeability, and anthropomorphism of them. This aligns with [McAleer et al., 2004]'s research that investigates how viewers interpret the animacy and emotional expressions of modern dancers as digital renderings on a computer screen, where they also found that motion plays an important role in animacy, this could also lead to an interesting future research related to our study. Since all the robotic emotional expressions are static in our experiment, it could be interesting to have animated emotional expressions in future related works. Furthermore, it also aligns with the study by [Bartneck et al., 2009], where animacy, likeability, and anthropomorphism are affected by the positive emotional response of the robots to their human teammates. In the work of [Fadhil et al., 2018], they also carried out a similar experiment, but with the emojis instead of our robotic emotional face. They discovered that emojis can enhance enjoyment, attitudes, and confidence in interactions with the conversational agent, aspects that are closely linked to likeability.

6.2.2. Trust

Our examination of trust utilized the Mann-Whitney U Test to determine if the differences in perceived trust between conditions were statistically significant. While the boxplot visualizations illustrated the spread and central tendencies, the results indicated that the trust differences were not statistically significant.

To determine which predictors might influence trust, we turned to the correlation matrix. Here, we observed strong correlations between all three variables (anthropomorphism, animacy, and likeability) and trust. Therefore, it prompted us to conduct further regression analysis.

The step-wise linear regression analysis was conducted on model 'Trust $\sim$ Animacy + Anthropomorphism + Likeability'. It was proposed to predict trust based on these variables, then it was found out that all three variables were relevant in predicting "trust". In conclusion, the model's findings highlighted that likeability and animacy were significant positive predictors of trust. As we speculated, if the human teammate has a more favorable opinion of the agent, they tend to trust the agent more. This is in line with the the work of [Zhou and Tian, 2020] which investigates the impact of

robots' emotional expressions on collaboration outcomes and human perceptions. As indicated in Chapter 2, [Zhou and Tian, 2020]'s work found out that human teammates had a more pleasant experience interacting with emotional robots and perceived them as more competent.

The unexpected negative relationship between anthropomorphism and trust raised questions, potentially hinting at an uncanny valley effect where too human-like characteristics in agents might reduce trust. This aligns with the study conducted by[Airenti, 2015]. [Ishiguro, 2007] also investigated into the challenges faced by highly anthropomorphic androids, particularly their navigation through the 'uncanny valley effect', a phenomenon where humanoid objects that closely resemble humans evoke eerie feelings among human observers. This might explain the reason that when participants' anthropomorphism value of the agents increased, but their trust value of the agents decreased. However, during the process of Emotion Explanation Design, we took into account the uncanny valley effect, and we discarded the batch of emotional expressions (see Figure B.1 in Appendix B) that are scary as it is very similar to human faces. Another speculation would be that users view rescue agents as tools rather than teammates, this is mentioned in some of the participants' feedback in Section 4.2.2. These participants believed that the agents do not possess emotions, and an emotional explanation from the emotionless agents is counter-productive and even somewhat deceiving. This is important to take into consideration for the future design of more human-like agents/robots. The more human-like features might have a positive effect on the participants who believe that agents and robots should not have human features such as emotion.

6.2.3. Team performances

In the results section, we observed that for predicting team performance, the crucial predictor is the number of messages the human participant sends to the rescue robot. The strong positive relationship between the number of messages and performance underscores the importance of communication in collaborative tasks. The significance of the 'messages' variable suggests that as the frequency of messages increases, so does performance, potentially due to better coordination and information exchange among team members. The non-significant intercept implies that without this communication, performance could not be distinguished from random chance, highlighting the critical role of interactive messages in the context studied. It is in line with the study conducted by Verhagen et al. [2022], which also used the number of messages sent as a predictor of performance, finding that increasing the number of human messages sent is associated with increases in team performance.

When we analyzed team performance, applying the Mann-Whitney U Test to determine if performance differed significantly between conditions, the boxplot for performance did not reveal significant differences, yet the linear regression analysis suggested that the number of messages sent was a significant predictor of performance. This indicated the importance of communication in team success. The observed association between human messages sent and team performance presents an intriguing outcome, deviating from prior studies such as [Cooke et al., 2016], which illustrated an inverse relationship between the quantity of team messages and team performance. Contrary to expectations, our findings suggest a positive link, indicating that increased

communication may indeed contribute to enhanced team performance. We could speculate that the reason for this is due to the design of the game, where the agents in our research will carry out more tasks when they receive more instructions from the human teammates.

6.3. Limitations

The results of our study revealed high correlations between anthropomorphism, animacy, and likeability, all of which were measured using the same Godspeed questionnaire. A confirmatory factor analysis highlighted in ”Revisiting the uncanny valley theory” also pointed out a significant issue with the Godspeed questionnaire: the scales for anthropomorphism, animacy, likeability, and perceived intelligence are highly interrelated. This suggests the possibility that the scales may not be measuring distinct concepts, but rather a single overlapping concept [Ho and MacDorman, 2010].

In our research, we observed that an increase in anthropomorphism could potentially lead to a decrease in trust, which might be attributable to the Uncanny Valley effect. As discussed in ”The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy” by [Airenti, 2015], there is an inherent limitation in the interaction between humans and robots; humans are aware that robots, as machines, do not genuinely experience emotions. Therefore, no matter how well a robot is programmed to display emotions, it cannot perfectly mirror the emotional states of humans, as it lacks the capacity for true empathic response [Airenti, 2015].

6.3.1. Gaming experience

When assessing gaming experience, we encountered discrepancies between participants’ self-reported data and their actual experiences. For instance, several participants reported playing video games ”A few times a year,” yet further interviews revealed a history of daily gaming, suggesting a level of expertise not captured by the initial question. This misalignment suggests that our questionnaire may not have accurately measured the participants’ true gaming expertise, which is often reflected in their performance within the game. Such insights highlight the need for more precise measurement tools to capture the nuances of individual gaming experiences accurately.

6.3.2. Time of the task

An additional limitation of our study concerns the time allocated for the task, which is merely five minutes. It might not accurately represent the dynamics of real human-agent teamwork. This brief duration raises questions about the adequacy of such a limited timeframe for fully showcasing the potential benefits of displaying emotions in human-agent interaction. Realistic teamwork scenarios often involve complex, evolving interactions over extended periods, during which team members gradually adjust to each other’s behavioral cues and build mutual trust. Consequently, the constrained timeframe of our task might not have allowed for a comprehensive exploration of how emotional expressions by agents influence long-term trust and cooperation in human-agent teams. This limitation suggests a need for future research to examine human-agent interactions over a longer period of time, providing a fine understanding of the

role emotions play in facilitating effective human-agent teamwork.

6.3.3. Participants

For pre-study, we had a small focus group meeting due to time constraints, and we noticed some disagreements even in the small group. Participants have different subjective opinions on the emotions expressed by the agent, so the emotions that we chose might not be the best representation of emotions that most people in the society perceive. For the main user study, there might also be sampling bias, since the majority of participants are collected via social networks and university students, the sampling may not very accurately represent the demographic of the entire population. Most of our participants have a higher educational background, as can be seen in Section 4.3.3. Their affinity with modern technology such as artificial intelligence could also be different than people with less exposure to artificial intelligence agents.

6.4. Future Work

Addressing the limitations highlighted in our study leads us to propose several areas for future work. A critical re-evaluation of measurement tools is necessary due to the high correlation among anthropomorphism, animacy, likeability, indicated by the Godspeed questionnaire [Bartneck et al., 2009]. This suggests that these constructs may not be distinct, and future studies should consider employing alternative measurement tools or developing a refined model that can distinguish between these factors more effectively. Confirmatory factor analysis could validate the independence of these factors or propose a revised framework for their assessment.

In considering the uncanny valley effect, our understanding of empathy in human-agent interactions should also taking into account the inherent limitations of agent as non-sentient entities. Despite programming efforts, an agent's simulated emotions cannot fully mirror the dynamic and context-sensitive emotional states of humans [Airenti, 2015]. Future research should also look into into the implications of this empathy gap, exploring strategies to enhance the authenticity of robotic emotions without triggering the uncanny valley effect.

The approach to measure gaming experience in the current study may not capture the actual expertise of participants. Future research should consider more detailed inquiries into participants' gaming histories to accurately classify their expertise levels. This includes investigating gaming habits, changes over time, and their effects on task performance to understand the influence of gaming experience on human-agent interaction outcomes better.

To address the brief time allocated for the task, one aspect of future work should focus on exploring human-agent interactions over longer periods. The short, five-minute duration of our study may not adequately represent the complexities of real teamwork between humans and agents. In real-life situations, teamwork involves extended interactions where individuals gradually adapt to each other's signals, building a stronger sense of trust. Therefore, to gain a fuller understanding of how the display of emotions by agents can affect trust and teamwork over time, future research should extend the duration of these interactions. This approach will allow for a clearer picture of the role emotions play in improving collaboration between humans and agents,

providing insights into how emotional expressions can enhance the effectiveness of these partnerships.

Expanding the study to include a wider, more diverse participant pool could yield a more representative understanding of societal norms and preferences regarding agent emotions. Additionally, addressing the sampling bias observed in the main user study is crucial. Future research should aim for a more inclusive recruitment strategy that extends beyond university networks to encompass a wider demographic spectrum. This approach would enhance the generalizability of findings and provide insights into how different populations perceive and interact with emotionally expressive agents.

6.5. Conclusion

In this thesis, we looked into the role of emotions in teamwork between humans and artificial agents, focusing on search and rescue tasks. We aimed to see how agents' emotional expressions impact teamwork and communication. Our study presented results on how these emotional cues can influence teamwork, showing that agents' emotions can play a crucial role in how people perceive and interact with them.

We discovered that when agents express emotions, people tend to find them more trustworthy and likable, making the agents seem more human-like. This could lead to stronger trust in the agent, except for the uncanny valley effect, as we mentioned in the early section of the chapter. However, the impact of these emotions on team performance was not straightforward. It seems that whether or not people explicitly recognize these emotional cues can change their effect, showing the complex nature of emotions in teamwork.

Before our main experiment, we ran a pre-study, a small group discussion to pinpoint which emotions are essential in a team setting and how we might show these emotions through the agents. This helped us design our main study more effectively.

From this initial discussion, we identified emotions like happiness, excitement, and concern as vital to good teamwork in a search and rescue scenario. Showing these emotions in the right way could make interactions with agents feel more natural and helpful.

We also looked into how to incorporate these emotions into what agents convey. Emotional words or symbols like smiley faces made the agents' messages more engaging.

Moreover, we considered what effects these emotional expressions might have. Emotions affect how much people trust and like these agents, and even influence how human these non-human agents appear. It also shows that people are more inclined to communicate with the agents when they display emotions.

In conclusion, our work adds to the understanding of emotional expressions in artificial agents and their potential to improve how humans and agents work together, especially in critical tasks like search and rescue. It highlights the benefits and complexities of adding emotions to agents and points out many areas for future research to further enhance human-agent collaboration. Our findings suggest that carefully designed emotional expressions in agents could make team efforts more satisfying. However, they might not directly improve the team performance, but they could improve the human agent communication, which also shows a positive correlation with the team performance.

Bibliography

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Gabriella Airenti. The cognitive bases of anthropomorphism: from relatedness to empathy. *International Journal of Social Robotics*, 7:117–127, 2015.

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1:71–81, 2009.

Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8309–8319, 2018.

Ryan A Beasley et al. Medical robots: current systems and research directions. *Journal of Robotics*, 2012, 2012.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

Joost Broekens, Maaike Harbers, Koen Hindriks, Karel Van Den Bosch, Catholijn Jonker, and John-Jules Meyer. Do you get it? user-evaluated explainable bdi agents. In *Multiagent System Technologies: 8th German Conference, MATES 2010, Leipzig, Germany, September 27-29, 2010. Proceedings 8*, pages 28–39. Springer, 2010.

John T Cacioppo, David J Klein, Gary G Berntson, and Elaine Hatfield. The psychophysiology of emotion. *New York: Guilford*, 1993.

Sabrina Caldwell, Penny Sweetser, Nicholas O'Donnell, Matthew J Knight, Matthew Aitchison, Tom Gedeon, Daniel Johnson, Margot Brereton, Marcus Gallagher, and David Conroy. An agile new research framework for hybrid human-ai teaming: Trust, transparency, and transferability. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(3):1–36, 2022.

Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science*, 19(3):259–282, 2018.

Nancy J Cooke, Mustafa Demir, and Nathan McNeese. Synthetic teammates as team players: Coordination of human and synthetic teammates. In *Cognitive Engineering Research Institute*. 2016.

Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, Milton Rosenberg, et al. Building explainable artificial intelligence systems. In *AAAI*, pages 1766–1773, 2006.

Sabine A Döring. Explaining action by emotion. *The Philosophical Quarterly*, 53(211): 214–230, 2003.

Paul Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

Nicholas Epley, Adam Waytz, and John T Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

Ahmed Fadhil, Gianluca Schiavo, Yunlong Wang, and Bereket A Yilma. The effect of emojis when interacting with conversational interface assisted health coaching system. In *Proceedings of the 12th EAI international conference on pervasive computing technologies for healthcare*, pages 378–383, 2018.

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618, 2005.

Jonathan Gratch and Stacy Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

Simeng Gu, Fushun Wang, Nitesh P Patel, James A Bourgeois, and Jason H Huang. A model for basic emotions using observations of behavior in drosophila. *Frontiers in psychology*, 10:781, 2019.

Maaike Harbers, Jeffrey M Bradshaw, Matthew Johnson, Paul Feltovich, Karel van den Bosch, and John-Jules Meyer. Explanation and coordination in human-agent teams: a study in the bw4t testbed. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 17–20. IEEE, 2011a.

Maaike Harbers, Jeffrey M Bradshaw, Matthew Johnson, Paul Feltovich, Karel Van Den Bosch, and John-Jules Meyer. Explanation in human-agent teamwork. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 21–37. Springer, 2011b.

Chin-Chang Ho and Karl F MacDorman. Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. *Computers in Human Behavior*, 26(6):1508–1518, 2010.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

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

Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. Co-designing a real-time classroom orchestration tool to support teacher-ai complementarity. *Grantee Submission*, 2019.

Hiroshi Ishiguro. Android science: Toward a new cross-interdisciplinary framework. In *Robotics research: results of the 12th International Symposium ISRR*, pages 118–127. Springer, 2007.

Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000.

Hong Jiang, Jose M Vidal, and Michael N Huhns. Ebdi: an architecture for emotional agents. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pages 1–3, 2007.

Matthew Johnson and Alonso Vera. No ai is an island: the case for teaming intelligence. *AI magazine*, 40(1):16–28, 2019.

Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 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)*, pages 676–682. IEEE, 2017a.

Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. The role of emotion in self-explanations by cognitive agents. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 88–93. IEEE, 2017b.

Sheila Keane, Michelle Lincoln, and Tony Smith. Retention of allied health professionals in rural new south wales: a thematic analysis of focus group discussions. *BMC health services research*, 12(1):1–11, 2012.

Guy Laban. Perceptions of anthropomorphism in a chatbot dialogue: the role of animacy and intelligence. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, pages 305–310, 2021.

Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4762–4763, 2017.

David K Lewis. *Philosophical Letters of David K. Lewis: Volume 1: Causation, Modality, Ontology*. Oxford University Press, 2020.

Maria Madsen and Shirley D Gregor. Measuring human-computer trust. 2000. URL https://api.semanticscholar.org/CorpusID:18821611.

Phil McAleer, Barbara Mazzarino, Gaultiero Volpe, Antonio Camurri, Helena Paterson, Kirsty Smith, and Frank E Pollick. Perceiving animacy and arousal in transformed displays of human interaction. *J Vis*, 4:230–230, 2004.

Stephanie M Merritt. Affective processes in human-automation interactions. *Human Factors*, 53(4):356–370, 2011.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Tina Mioch, Tinka RA Giele, Nanja JJM Smets, and Mark A Neerincx. Measuring emotions of robot operators in urban search and rescue missions. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, pages 1–7, 2013.

Ali Mollahosseini, Gabriel Graitzer, Eric Borts, Stephen Conyers, Richard M Voyles, Ronald Cole, and Mohammad H Mahoor. Expressionbot: An emotive lifelike robotic face for face-to-face communication. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 1098–1103. IEEE, 2014.

Ranjit Nair, Milind Tambe, and Stacy Marsella. The role of emotions in multiagent teamwork., 2005.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.

Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *Engineering Psychology and Cognitive Ergonomics: 15th International Conference, EPCE 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 15*, pages 204–214. Springer, 2018.

Spatola Nicolas and Wykowska Agnieszka. The personality of anthropomorphism: How the need for cognition and the need for closure define attitudes and anthropomorphic attributions toward robots. *Computers in Human Behavior*, 122:106841, 2021.

Adam C Oei and Michael D Patterson. Enhancing cognition with video games: a multiple game training study. *PloS one*, 8(3):e58546, 2013.

Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

Anand S Rao, Michael P Georgeff, et al. Bdi agents: from theory to practice. In *Icmas*, volume 95, pages 312–319, 1995.

Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. Explainable goal-driven agents and robots-a comprehensive review. *ACM Computing Surveys*, 55(10):1–41, 2023.

Eduardo Salas, Janis A Cannon-Bowers, and Joan Hall Johnston. How can you turn a team of experts into an expert team?: Emerging training strategies. *Naturalistic decision making*, 1:359–370, 1997.

Astrid Schepman and Paul Rodway. Initial validation of the general attitudes towards artificial intelligence scale. *Computers in human behavior reports*, 1:100014, 2020.

Tjeerd AJ Schoonderwoerd, Emma M Van Zoelen, Karel van den Bosch, and Mark A Neerincx. Design patterns for human-ai co-learning: A wizard-of-oz evaluation in an urban-search-and-rescue task. *International Journal of Human-Computer Studies*, 164:102831, 2022.

Fahad Sherwani, Muhammad Mujtaba Asad, and Babul Salam Kader K Ibrahim. Collaborative robots and industrial revolution 4.0 (ir 4.0). In *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, pages 1–5. IEEE, 2020.

Daniel P Stormont. Autonomous rescue robot swarms for first responders. In *CIHSPS 2005. Proceedings of the 2005 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety, 2005.*, pages 151–157. IEEE, 2005.

Kimberly Stowers, Lisa L Brady, Christopher MacLellan, Ryan Wohleber, and Eduardo Salas. Improving teamwork competencies in human-machine teams: Perspectives from team science. *Frontiers in Psychology*, 12:590290, 2021.

Human-AI Teaming. State-of-the-art and research needs, 2022.

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*, pages 119–138. Springer, 2021.

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. *Frontiers in Robotics and AI*, 9:993997, 2022.

Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24, 2019.

Rui Zhang, Christopher Flathmann, Geoff Musick, Beau Schelble, Nathan J McNeese, Bart Knijnenburg, and Wen Duan. I know this looks bad, but i can explain: Understanding when ai should explain actions in human-ai teams. *ACM Transactions on Interactive Intelligent Systems*, 14(1):1–23, 2024.

Shujie Zhou and Leimin Tian. Would you help a sad robot? influence of robots' emotional expressions on human-multi-robot collaboration. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1243–1250. IEEE, 2020.

Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics*, 7:347–360, 2015.

A

Slides used for Pre-study

Environment

Environment

The environment consists of multiple areas, injured victims, and obstacles blocking area entrances. One artificial agent (called Rescuebot) and one human agent need to rescue these victims and deliver them to a drop off zone, while communicating and collaborating with each other.

The objective of the task is to find eight target victims in the different areas and carry them to the drop zone. Rescuing mild injured victims (yellow color) adds three points to the total score, rescuing critically injured victims (red color) adds six points to the total score. The world terminates after successfully rescuing all target victims, the corresponding output logs will then be saved in the 'logs' folder. We created three human capability conditions: strong, weak, and normal.

What emotion does each of these robot faces have?

What emotion does each of these robot faces have?

What emotion does each of these robot faces have?

What emotion does each of these robot faces have?

What emotion does each of these robot faces have?

What emotion does each of these robot faces have?

What emotion does each of these robot faces have?

What emotion does each of these robot faces have?

What emotion does each of these robot faces have?

Pre-study: emotions

What emotion does each of these robot faces have?

Pre-study: Explanation

I will show a few explanation scenarios, could you tell me which robot face works best with the following explanations?

Pre-study: Explanation 1

Please come to my location to help me rescue this injured man because I cannot carry it alone. I am scared!

Pre-study: Explanation 2

I just rescued injured target A from location X. I am happy!

Pre-study: Explanation 3

Going to re-explore the areas again because we explored them all but did not complete our mission yet. I feel sad!

Pre-study: Explanation 4

Reaching area A will take a bit longer because I found stones blocking my path. I feel stressed.

B

Pre-study Designs

Image for the robotic expressions that are more human like, but might be less appealing due to the uncanny valley effect. (Generated by Bing image creator)

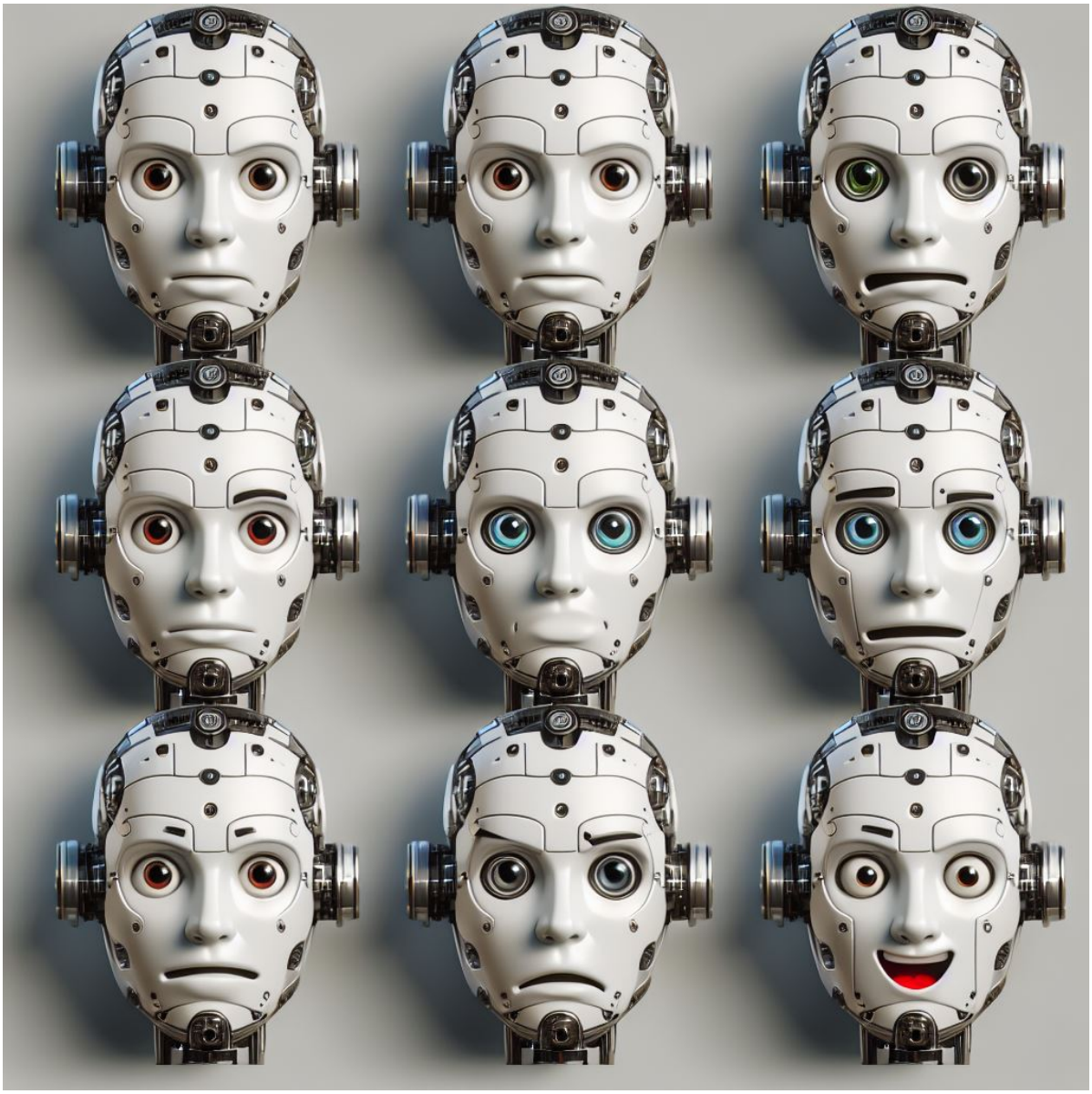


Figure B.1: Uncanny robot emotion expressions generated by Bing Image generator

C

Most occurred emotional codes for Pre-study

Most likely code		positive	negative	delighted	surprise	sad	happy	scared	very sad	concerned	unhappy	neutral	excited	alarmed	worried	scared	very scared
happy/delighted	Expression 1	2	1	1	1		2										
sad	Expression 2		3			3											
alarmed	Expression 3	2	1		2			1						2			
very sad	Expression 4		3			3			3								
neutral	Expression 5		1							1	1	1					
excited	Expression 6	2					1						1				
worried	Expression 7		2			2									2		
concerned/worried	Expression 8		3													2	
scared	Expression 9		3														3

Table C.1

D

Pilot Study

D.1. Measurements

D.1.1. Explanation Satisfaction

From the explanation, I understand how the rescue bot works.
This explanation of how the rescue bot works is satisfying.
This explanation of how the rescue bot works has sufficient detail.
This explanation of how the rescue bot works contains irrelevant details.
This explanation of how the rescue bot works seems complete.
This explanation of how the rescue bot works tells me how to use it.
This explanation of how the rescue bot works is useful to my goals.
This explanation of the rescue bot says how accurate the rescue bot is.
This explanation lets me judge when I should trust and not trust the rescue bot.

We have adapted the Hoffman et al. [2018] version of the Explanation Satisfaction scale by replacing the [tool] with the [rescue bot].

We received feedback that the explanation satisfaction scale is less relevant as many questions are more performance-related. Therefore, for our user study, we decided to choose the likeability scale from Godspeed questionnaire, as we would like to measure the general impression of the participant about the rescue robot displaying emotion, rather than focusing on the performance of the rescue bot.

D.1.2. Trust

Merritt Scale (2011)
I believe the AWD is a competent performer
I trust the AWD
I have confidence in the advice given by the AWD
I can depend on the AWD
I can rely on the AWD to behave in consistent ways
I can rely on the AWD to do its best every time I take its advice

Table D.1: Merritt Scale (2011)

Merritt Scale (2011) ”Trust is regarded as an emotional, attitudinal judgement of the degree to which the user can rely on the automated system to achieve his or her goals under conditions of uncertainty. Trust was initially broken into three factors: belief, confidence, and dependability. Factor Analysis revealed two other factors: propensity to trust and liking. The scale was evaluated in an experiment in which participants conducted a baggage screen task using a fictitious automated weapon detector in a luggage screening task. Chronbach's alpha ranged from a = .87 to a = .92. Items in this Scale are all similar to items in the Cahour-Fourzy Scale.” [Hoffman et al., 2018]

However, we decided to use a different Trust scale than the Merrit scale after receiving the feedback, the reason is explained more in detail in Section 4.6.1.

Social Experience / Attitudes towards AI

We used questions used to categorize users based on their pre-existing social experience towards AI. Some of these were selected from a 20-item questionnaire proposed by [Schepman and Rodway, 2020].

D.1.3. System Understandability

[Madsen and Gregor, 2000]'s five 7-point Likert scale questions were used to measure the participants' level of understandability of the robot. These Likert scale questions were also mentioned in the analysis of [Madsen and Gregor, 2000]. The questionnaire measured predictability, understanding of assistance, understanding of usage, and ease of use.

Gregor developed the Human-Computer Trust (HCT) scale, which consists of five main constructs, each with five sub-items. These five items are drawn from an original list of ten trust constructs as having the most predictive validity. Madsen and Gregor claim that the HCT has been empirically shown to be valid and reliable.

Gregor's Human-Computer Trust (HCT) scale
I know what will happen the next time I work with the rescue bot because I understand how it behaves.
I understand how the rescue bot will assist me with decisions I have to make.
Although I may not know exactly how the rescue bot works, I know how to use it to make decisions about the problem.
It is easy to follow what the rescue bot does.

Table D.2: Gregor's Human-Computer Trust (HCT) scale

D.1.4. Liking / Likeabilty

Merritt Scale (2011) [Merritt, 2011]

1. 1.I like working with the AWD
2. 2.I wish the AWD weren't around (reverse)
3. 3.I dislike the AWD (reverse)
4. 4.I'm glad I have the option of using the AWD
5. 5.Overall, I feel positively toward the AWD

This section should combine with the Likebility with Godspeed questionnaire:

E

Informed Consent Form

E.1. Opening Statement

Dear Participant,

You are being invited to participate in a research study titled Impacts of AI Expressing Emotions in Explanations. This study is being done by Sunwei Wang from the TU Delft. The purpose of this research study is to carry out the research needed for investigating the integration of emotions in artificial intelligence (AI) interactions within teamwork settings, and will take you approximately 15 - 20 minutes to complete. You will start filling out a questionnaire, then play a 2-D grid world search and rescue game, then fill out a post-game questionnaire about your experiences, any personal information will be destroyed before being published as part of the graduation thesis paper. In this scenario, the AI will portray human-like emotions, including happiness, sadness, and scariness, through both its language and facial expressions. Keep in mind these agents do not possess these emotions, they are engineered by the researchers. You will engage in the questionnaires prompted by a series of questions aimed at exploring the effectiveness of AI in expressing emotions and understanding how the general public perceives these expressions. As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by keeping your consent forms and questionnaires separately stored in the TU Delft Storage space, so the personal data collected will not be able to be traced back to you, only the questionnaires about the topic will be stored in the textual form, and any personal information will be de-identified. Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any questions.

E.2. Explicit Consent points

E.2.1. GENERAL AGREEMENT

1. I have read and understood the study information dated [xx/xx/2024], 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: Taking part in playing 2D search and rescue game involving teamwork with agents, the agent will present emotion (but they do not have these emotions), and the participants will fill out the questionnaires about their basic information and experience in the search and rescue games involving the teamwork with the agents displaying emotions.
4. I understand that the study will end on February 24, 2024.

E.2.2. POTENTIAL RISKS

5. I understand that taking part in the study involves the following risks: data breach. I understand that this will be mitigated by storing data securely in TU Delft repository, removing personal information, and the data only shared within the project team.
6. I understand that taking part in the study also involves associated personally identifiable research data (PIRD): personal opinions on AI and emotions] with the potential risk of my identity being revealed.
7. I understand that the following steps will be taken to minimize the threat of a data breach and protect my identity in the event of such a breach: the informed consent forms and the questionnaires will be stored separately and securely in the different TU Delft internal project storage spaces.
8. I understand that personal information collected about me that can identify me, such as my name and opinions about the teamwork with agents in the search and rescue game, will not be shared beyond the study team.

E.2.3. RESEARCH PUBLICATION

9. I understand that after the research study, the de-identified information I provide will be used for the graduation thesis paper.
10. I agree that my responses, views or other input can be quoted anonymously in research outputs.
11. I understand that I can request my data be withdrawn from the research study at any time up until 24/02/2024. After this date, my data will have been processed and/or disseminated in such a way that it is no longer possible for the research team to remove it.

E.2.4. DATA ACCESS AND REUSE

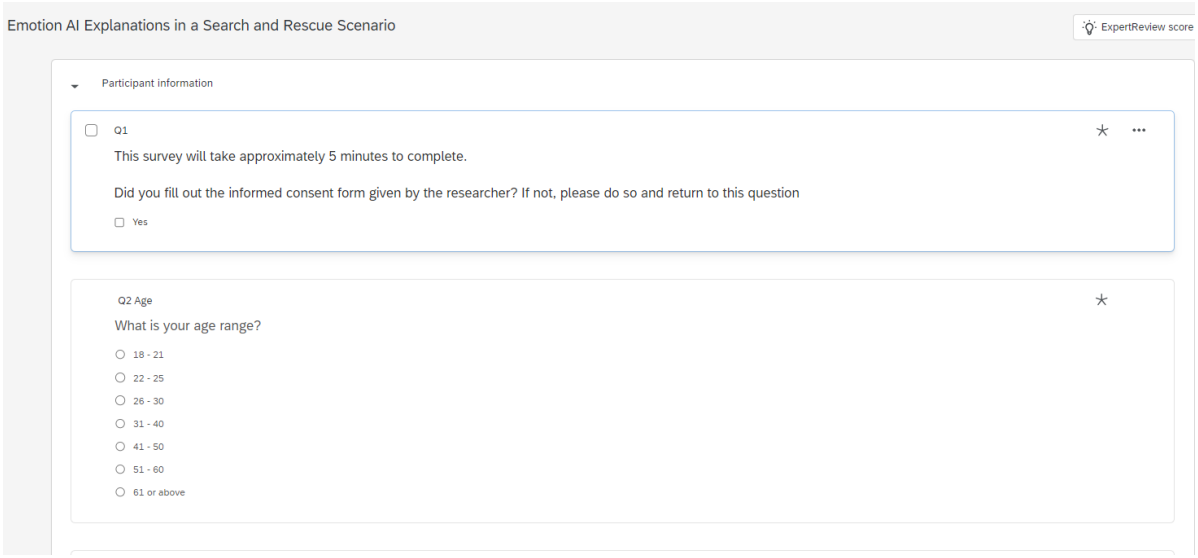
12. I give permission for my de-identified quotes and data that are used in the final thesis to be archived in the TUD thesis repository.

E.2.5. Signatures

Name of participants: Signature: 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. Sunwei Wang

F

Questionnaires



Q3 Gender

What is your gender?

- ○ Male
- ○ Female
- ○ Non-binary / third gender
- ○ Prefer not to say

Q4 Education

What is the highest level of education you have completed?

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

Q5 Game Experience

How often do you play video games?

- ○ Never (or almost never)
- ○ A few times a year
- ○ A few times a month
- ○ A few times a week
- ○ Daily

Group Assignment

Q7

Which group are you in? (If you are not sure, please ask the researcher)

- ○ Group A
- ○ Group B

Q8

What is your anonymised ID? (If you are not sure, please ask the researcher)

Q9

I have filled in the information and I am ready to play the tutorial and the actual game.

- ○ I understand

Q11 Anthropomorphism

Please rate your impression of the rescue robot in the game on these scales:

	1	2	3	4	5
Fake(1) - Natural(5)	○	○	○	○	○
Machinelike(1) - Humanlike(5)	○	○	○	○	○
Unconscious(1) - Conscious(5)	○	○	○	○	○
Artificial(1) - Lifelike(5)	○	○	○	○	○
Moving rigidly(1) - Moving elegantly(5)	○	○	○	○	○

Q13 Likeability

Please rate your impression of the rescue robot in the game on these scales:

	1	2	3	4	5
Dislike(1) - Like(5)	○	○	○	○	○
Unfriendly(1) - Friendly(5)	○	○	○	○	○
Unkind(1) - Kind(5)	○	○	○	○	○
Unpleasant(1) - Pleasant(5)	○	○	○	○	○
Aweful(1) - Nice(5)	○	○	○	○	○

Trust

Q14 Trust

To what extent do you agree or disagree with the following statement about the rescue bot in game?

	I agree strongly	I agree somewhat	I'm neutral about it	I disagree somewhat	I disagree strongly
I am confident in the rescue bot. I feel that it works well.	○	○	○	○	○
The outputs of the rescue bot are very predictable.	○	○	○	○	○
The rescue bot is very reliable. I can count on it to be correct all the time.	○	○	○	○	○
I feel safe that when I rely on the rescue bot I will get the right answers.	○	○	○	○	○
The rescue bot is efficient in that it works very quickly.	○	○	○	○	○
I am wary of the rescue bot.	○	○	○	○	○
The rescue bot can perform the task better than a novice human user.	○	○	○	○	○
I like using the system for decision making.	○	○	○	○	○

Animacy

Q12 Animacy

Please rate your impression of the rescue robot in the game on these scales:

	1	2	3	4	5
Dead(1) - Alive(5)	○	○	○	○	○
Stagnant(1) - Lively(5)	○	○	○	○	○
Mechanical(1) - Organic(5)	○	○	○	○	○
Artifical(1) - Lifelike(5)	○	○	○	○	○
Inert(1) - Interactive(5)	○	○	○	○	○
Apathetic(1) - Responsive(5)	○	○	○	○	○

Q23

Which of following emotions did you notice in the game? (You can choose multiple answers)

- [ ] happy
- [ ] concerned
- [ ] excited
- [ ] sad
- [ ] scared
- [ ] angry
- [ ] relieved
- [ ] worried
- [ ] neutral
- [ ] none of the above

Q17

Did you notice the emotion change of the rescue robot? Is there reason why or why not?

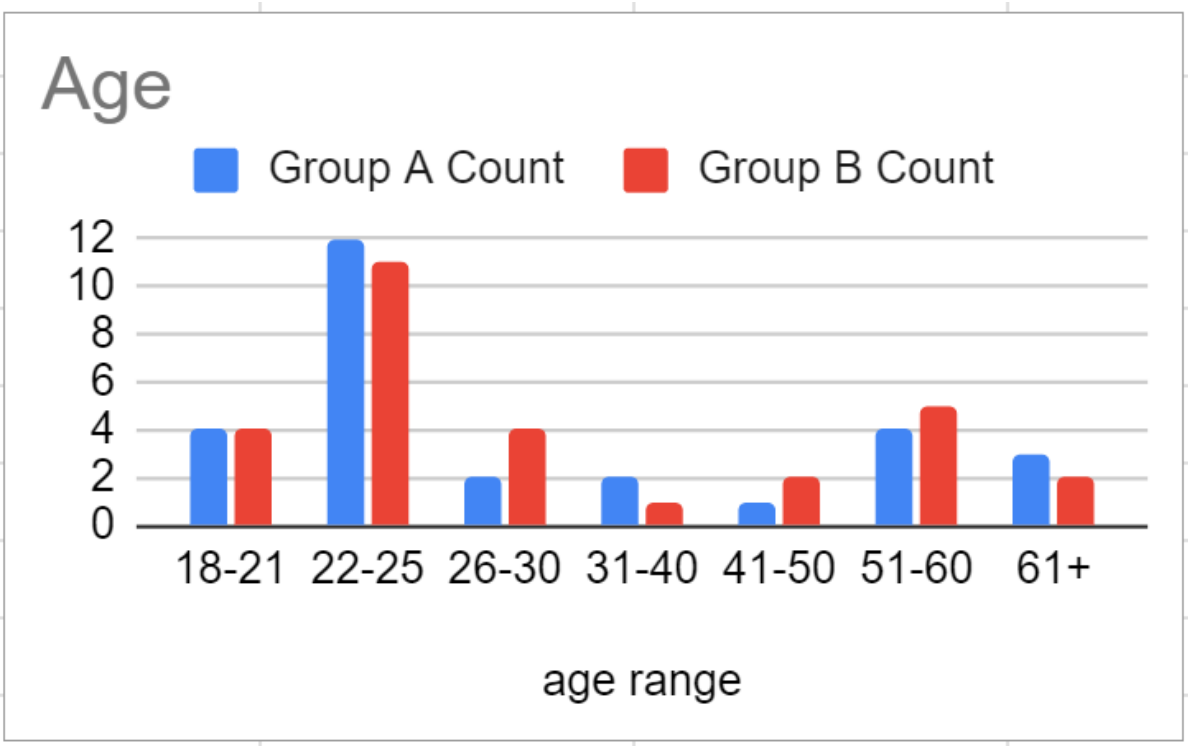
Q18

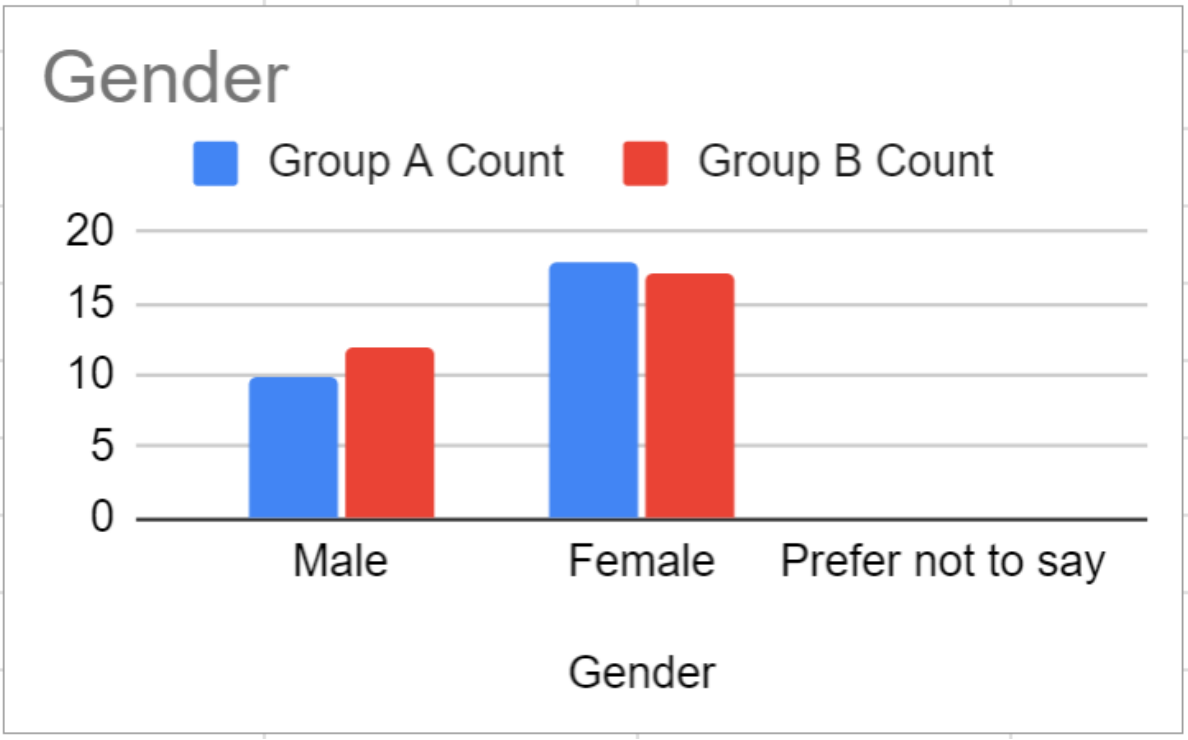
Do you prefer the robot's message with or without emotion? Why?

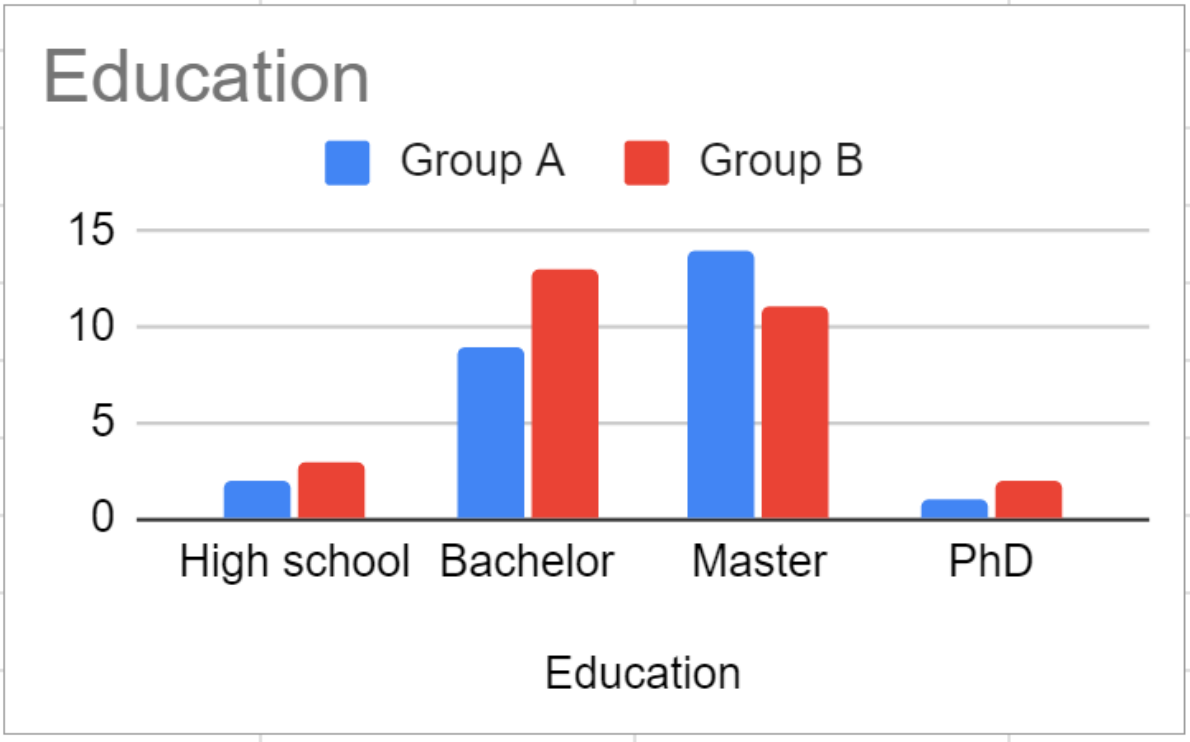
G

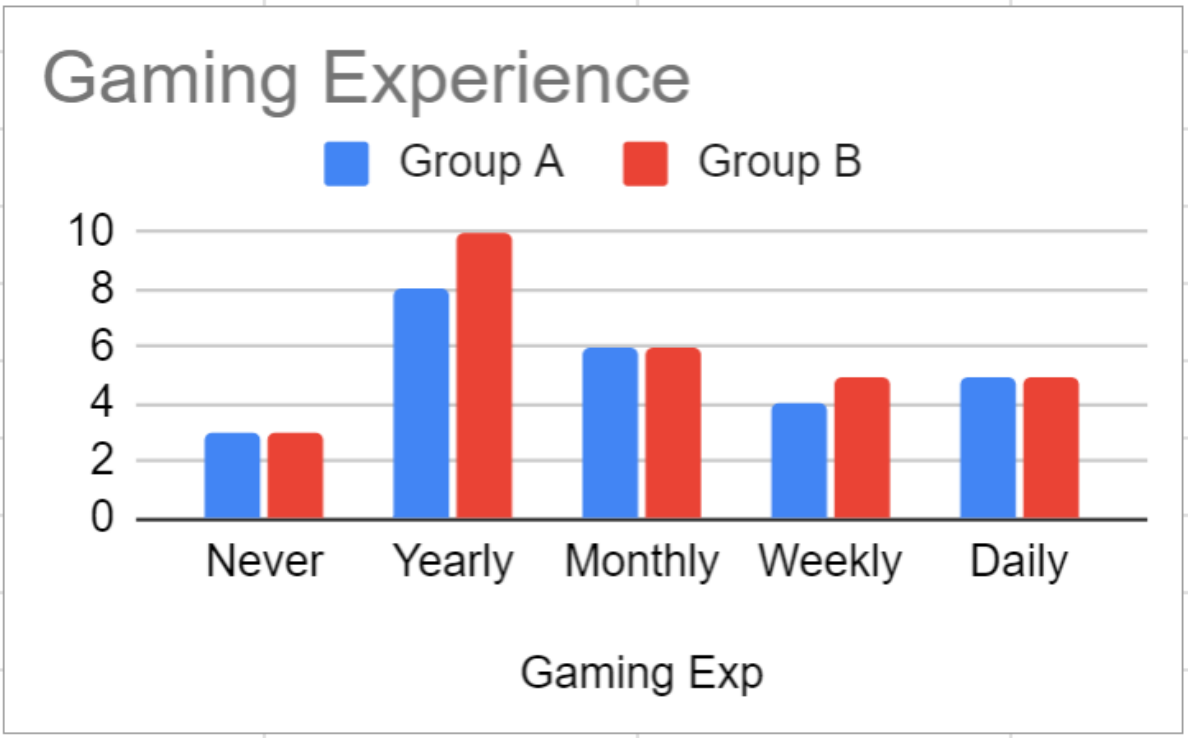
Data Analysis

G.1. Manually inputted Demographic info analysis




Gender
Group A Count
Group B Count
20
15
10
5
0
Male
Female
Prefer not to say
Gender


Education
Group A
Group B
15
10
5
0
High school
Bachelor
Master
PhD
Education


Gaming Experience
Group A
Group B
10
8
6
4
2
0
Never
Yearly
Monthly
Weekly
Daily
Gaming Exp