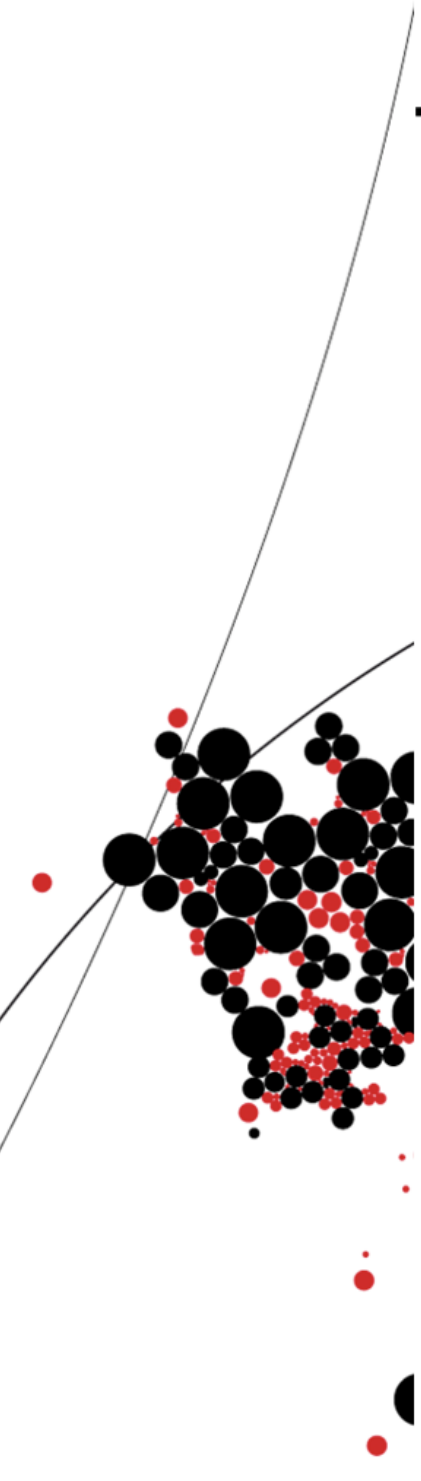
UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

The influence of anthropomorphism on our feelings for faulty robots

I.M.F. Bistolfi

M.Sc. Thesis
March 2022



Supervisors:
Dr. K.P. Truong
M.Sc. P.C. Velner
Ir. T.H.J. Beelen

University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Summary

Erroneous interactions are a frequent occurrence in Human Robot Interaction. For a robot to be successful and accepted by its user it is necessary to understand what influences the user's perception of the interaction. Error severity and level of anthropomorphism are both components of an interaction that heavily influence the user's perception. The perception of the user can be measured by looking at the User Experience of the interaction. More specifically, the trust and likability towards a robot are important when looking at the user's experience of an erroneous interaction. Therefore this research aimed to get an answer to the question: *To what extent do the appearance of a robot (high-anthropomorphic vs. low-anthropomorphic) and the error severity (high-severity vs. low-severity) influence the level of trust and likability of a robot in collaborative scenarios?*

To find an answer to this question a user study was carried out where 21 participants interacted with a robot in a virtual treasure hunting game. The users interacted with two robots with different levels of anthropomorphism. During these interactions the robots made identical errors. Error severity level was researched using a between-subjects study design, while level of anthropomorphism was researched using within-subjects design. The participants were asked to give an initial assessment of their likability and trust towards the robots before the game started.

The results showed that the level of anthropomorphism has a significant effect on the overall likability score of that robot, where high anthropomorphic robots have a significant positive effect on the overall likability score. However, when comparing the initial likability measurements with post-study measurements the growth/loss of likability due to the effect of the level of anthropomorphism was not significant. Additionally, no significant effects were found on the interaction of anthropomorphism and error severity on likability scores. Similarly, no significant effect was found of the level of error severity on the likability scores. Furthermore, a high level of anthropomorphism had a significantly positive effect on the overall trust score. However, a high level of anthropomorphism did not have a significantly positive effect on the trust growth/loss that was found when comparing the initial trust measurements with the post-study trust measurements. Moreover, for the interaction between anthropomorphism and error severity in terms of trust score growth a significant difference was

found, where it showed that high levels of anthropomorphism resulted in a higher level of trust growth in the case of mild errors. While in the case of severe errors the level of anthropomorphism did not have a significantly different effect on the trust growth/loss. Nevertheless, for the interaction between anthropomorphism and error severity on the overall trust score no significant effect was seen. Additionally, no significant effects were found when looking at the impact of error severity on trust.

Additional findings included, that some participants pointed to the different personalities of the robots as reasons for their preference towards one of the robots. However, the so called “personalities” of the robots were identical as their behavior was identical.

Concluding, it seems that anthropomorphism has an effect on both likability and trust. While error severity on its own had no impact of likability or trust. Additionally the combination of anthropomorphism and error severity has a significant impact on the trust growth, but not on overall trust scores or likability. These results should be held in light of the limitations of this study that show that the likability measurements left something to be desired.

Acknowledgement

This master thesis is the result of many months hard work. This thesis is the last step towards my master's degree in Interaction Technology. During the start of this process my supervisors and I had many brainstorm session about the possible subjects for my research and after considering many options we came to the conclusion that this project was the best fit for me.

I would like to take this opportunity to thank everyone who supported me during my thesis and my master studies in general. First of all I would like to express my gratitude towards my entire graduation committee. I want to thank my supervisors, Khiet Truong, Thomas Beelen and Ella Velner for their invaluable feedback and knowledge. They supported me throughout the whole process of my master thesis and were a great resource during this process. Many thanks to them for helping me find a project, creating time for us to have many meetings and helping me throughout these many months. I also want to thank my external supervisor Edwin Dertien for helping me out on such short notice with this project.

I am also grateful for Stef Klein Geltink, for his editing help, feedback, and moral support. Thanks should also go to all the different study participants, including friends, peers and other people, who were generous enough to help me with my research by participating in my study.

Contents

Chapter 1

Introduction

Social robots are continuously being used and updated to fit into the lives of people. An example of this is the use of social assistance and companion robots in elderly care [1]. However, these interactions between humans and robots do not always go without errors. For the usage of social robots to be truly successful, people need to accept robots and be willing to interact with them [2]. But when errors occur, people's views of robots can change. How people perceive robots when they make errors is a key component in understanding how to make social robots that people will accept. The user experience (UX) associated with the Human-Robot Interaction (HRI) has consequences for the acceptance of robots [3]. More specifically, the user's trust in the robot, and the likability associated with the robot are important for successful and pleasant interactions [4]–[6]. The trust towards a robot is directly related to how effective a robot is, the performance, and the use rate of the robot [7]. Similarly, the likability people feel towards a robot is associated with the intention of future use [8].

A lot of existing research concerning errors in robots focuses on human-like robots, with a lot of anthropomorphic qualities. The conclusions that are found in these studies only apply to those anthropomorphic robots. However, in the field of HRI, less anthropomorphic robots are also frequently used. Examples of these less anthropomorphic robots are Smart Speakers with voice assistants. Which are found in, among others, Google Home and Amazon Echo. For this reason, it is important to also look at the consequences errors have in less anthropomorphic robots. Specifically, the consequences errors have on the user's perception of a less anthropomorphic robot. When designing a robot it is important to know how that design affects the interaction. Therefore, knowing more about the influence of anthropomorphism in erroneous robots on the user's perception of the robot is very useful. Additionally, the different severity levels of an error can also have different impacts and are important to consider when looking at errors in (anthropomorphic) robots. More specifically, it is of importance to know whether an anthropomorphic design has the same influence on robots making low-severity errors as it has on robots

making high-severity errors. Since it seems that this is not always the case. An example of this is that in low severity error situations a human-like robot was understood better than a smart speaker, while in high severity error situations the smart speaker was understood better [9].

This research will present a study into the different effects that errors can have on the trust and likability of a robot with different levels of anthropomorphism. In this research, the level of severity of the errors will also be considered. This leads to the following research question:

> *To what extent do the appearance of a robot (high-anthropomorphic vs. low-anthropomorphic) and the error severity (high-severity vs. low-severity) influence the level of trust and likability towards a robot in collaborative scenarios?*

A study will be carried out to find answers to the research question. The study will use two different robots, a high-level anthropomorphic robot, and a low-level anthropomorphic robot. Each participant gets to participate in a treasure hunting game where the participant and the robot have to collaborate to win the game. The robot will either make a severe error or a minor error. The level of severity is based on the consequences of the error.

Chapter 2 of this paper includes relevant research on the context of errors, anthropomorphism, and the importance of user experience is explained. Additionally, in Chapter 3 a comprehensive overview of the study is presented. The results of the study are then displayed in Chapter 4, and a discussion and conclusion are given in Chapter 5.1.

Chapter 2

Related work

2.1 Errors in HRI

As discussed in the introduction, Chapter 1, errors are important in HRI. To understand how errors influence the user's perception of the interaction it is needed to have some background on errors.

When talking about errors, the words *error*, *failure*, and *fault* are sometimes used interchangeably while they have different meanings. *Failures* are defined by Brooks [10] as "a degraded state of ability which causes the behaviour or service being performed by the system to deviate from the ideal, normal, or correct functionality". Failures are caused by *errors*, which originate in the software or hardware of the robot. These errors are the result of one or more *faults* [10]–[12]. These faults can be anything that causes the system to go to the error state [11]. The distinction between an error and a failure is that the error happens in the context of the system (i.e. not being able to correctly identify a word) while the failure happens in the context of the service the robot provides (i.e. giving an incorrect response). Faults, errors, and failures are thus different aspects of the same error process. For the context of this research, failures are the most important part since they focus on the user's perception. However, error situations exists out of errors, failures and faults. The combination of errors, failures and faults is the error process as a whole. In this paper the complete error process will be taken into account and will be referred to as an error.

2.1.1 Error categorization

To correctly handle errors, it is important to look at the kind of error that is made. Every kind of error can have a different consequence on the user's perception of a robot after the error occurred [13]. Similarly, Honig et al. [14] found that not all categories of errors have similar results for the user experience. The classification

of errors can be done in a multitude of ways, according to either type, severity, or recoverability [11], [12], [15], [16].

Types

Some researchers group errors according to the behaviour of an error. For example, Gompei and Umemuro [17] differentiate between four different types of speech errors in their research: addition, drop, substitution, and swap errors. Here the classification is based on the failure that is made. A different approach is used by Woerdt and Haselager [18], who categorize two types of errors based on the reason for the error to occur. These two categories are: (1) due to lack of effort or (2) due to lack of ability of the robot. Their reason for looking at these two categories was to make the distinction between controllable and uncontrollable error results. They find that an error due to lack of ability is uncontrollable, e.g. dropping an object, while an error due to lack of effort, e.g. throwing an object, is controllable.

A different, and more common approach is classifying the types of errors by the context of the error. Giuliani et al. [15] identified two types of errors: (1) technical failures, which are defined by the fact that they are caused by technical faults of the robot, and (2) social norm violations, which are defined by that they deviate from normal social scripts or normal social signals. Other researchers used similar approaches: e.g. task-oriented errors vs. social errors [19], performance errors vs. social errors [20], processing errors vs. communication errors [10], technical failures vs. social cognitive errors, and technical failures vs. decision level errors [21].

Honig and Oron-Gilad [11] identify one step above social errors. They find two types of errors: technical failures and interaction failures. Interaction failures include social norm violations as well as human errors and environmental and other agents. What kind of failures fall under which type can be seen in Figure 2.1.

Severity

Severity is a categorization of errors and failures that is often used in research. Laprie [12] made a distinction between failures based on severity. He considered two failure modes: (1) *benign failures*, where the cost of the consequences of the failure does not outweigh the payout of the interaction if no failure had taken place. (2) *malign failures*, where the cost of the consequences outweighs the payout in the absence of failure.

Where some researchers adopt Laprie's approach [22], other researchers have a different manner of categorizing severity. A slightly different approach than that of Laprie was taken by Stiber and Huang [23] and Van Waveren et al. [24]. Stiber and Huang [23] determine the severity of a failure by the impact that the error has on the

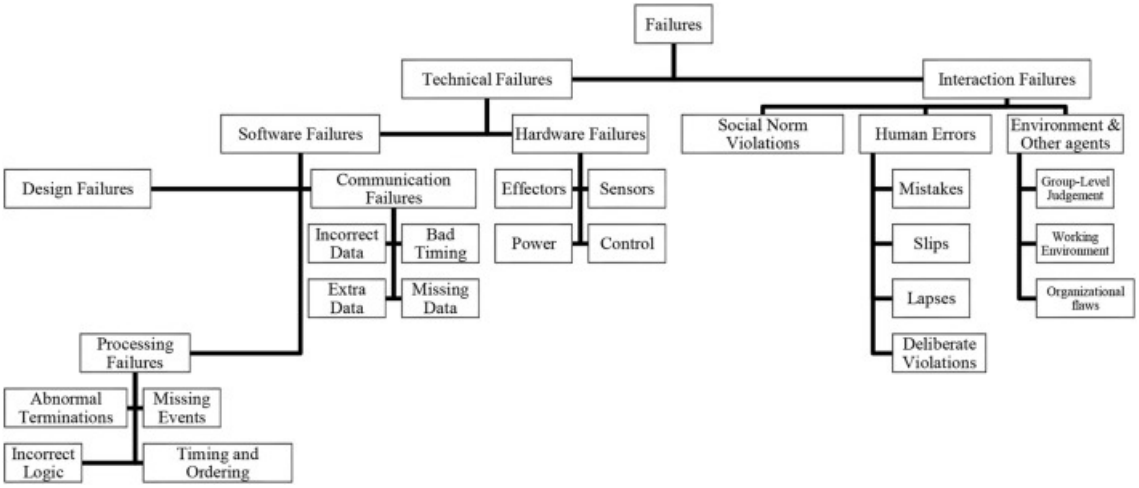


Figure 2.1: A tree overview of the categorization of different types of failures [11]

immediate surroundings. They categorize three levels of severity: low, medium, and high. While Van Waveren et al. [24] determine the severity of a failure by the impact that the error has on the user. Where the impact is defined by what is at stake for the user. They define two levels: low-impact and high-impact. Another approach to categorizing severity was found by Sarkar, Araiza-Illan, and Eder [25], who see an 'increase in faultiness' as an increase in severity, by adding two errors together they find the faultiness, and thus the severity of the error has increased. In other research time constraint is used as a manner of upping the severity level [9], [19].

When looking at error severity it is important to consider different contexts. Since the severity of the error is relative to the context in which the error takes place. For example, an error made in search and rescue robots can have significantly more severe consequences than an error made in entertainment robots. So it is important to note that a high severity error in one context can still be less severe than a high severity error in another context.

Recover-ability

Ross et al. [16] categorize errors on their ability to recover from the error. They define four different recover-abilities: *anticipated errors*: when the goal can still be achieved using a different course of action, *exceptional errors*: when the goal can still be achieved if a different plan is made since the current plan will not help to achieve the goal anymore, *unrecoverable errors*: when the current plan has failed and it has become impossible to achieve the goal, and *socially recoverable errors*: when the action has failed but the goal can still be achieved with the original plan with assistance from other agents in its environment. An important note here is that the recoverability described by Ross et al. [16] is focused on still achieving the goal

after a failure, rather than social recovery between the robot and the user. The latter is the focus of many pieces of research that look at the user's experience after a failure [26]–[28].

2.1.2 Error contexts

Error situations can lead to negative consequences in the user's perception of the robot. The context in which these errors happen can vary. A lot of research has been carried out into error situations in HRI in different contexts. Examples of contexts in which errors occur and are researched are: playing games with robots [29] [24], having a coworker assistant [25], companion robots in home environment [22] [30], assistance robots who guide tasks [9] [19] [31] [32] [33] [34], and learning companion robots [35]. A more detailed overview of these contexts can be found in Table 2.1. This table also shows that there are many different ways to approach errors in research.

Context	Scenario	Error	Source
Games	The participant uses voice commands to navigate the robot in a shooting game.	Unpredictable behaviour: the robot does something else than what is asked of it. Low functionality: the robot is unable to follow a command.	[29]
Games	The participant and robot work together to solve an escape room.	Low impact error: the robot would fail but still complete the escape room. High impact error: the robot would fail and they would lose the game.	[24]

Coworker assistant	The participant and the robot do a collaborative manufacturing task where they assemble a toy race car.	Three degrees of severity: 1. unable to pick up a component, but correctly instructs the participant. 2. gives wrong instructions for the initial steps of the assembly, but its motion and component picking up is correct. 3. both of the first two errors combined.	[25]
Home companion	An online interaction with a home companion robot in a virtual house.	Trivial errors: the robot makes a mistake that does not have big consequences. Severe errors: the robot makes a mistake that has big consequences.	[22]
Home companion	The participants were asked to interact with their friend's robotic home assistant that would welcome them.	Cognitive and physical imperfections on the robot by displaying incorrect memory and erratic movement.	[30]
Assistance with guided tasks	The participants were instructed by the robots on how to make a spring roll.	High severity: errors when there is time pressure. Low severity: errors when there is no time pressure.	[9]
Assistance with guided tasks	The participants were instructed by the robots on how to make a spring roll.	High severity: errors when there is time pressure. Low severity: errors when there is no time pressure. Types of errors: Disengagement, incomplete instruction, no response, repeating, incorrect guidance.	[19]

Assistance with guided tasks	The participant would make an omelette, the robot helps them by handing them the ingredients.	Dropping the egg.	[31]
Assistance with guided tasks	A LEGO building session, where the robot instructs the participant.	4 Social norm violations around the timing of speech and abnormal instructions. 4 Technical failures around speech errors and failing to do a task.	[32]
Assistance with guided tasks	Unpacking moving boxes, where the robot tells the participant where the items should go.	No gestures or incongruent gestures to accompany speech.	[33]
Assistance with guided tasks	The participant is asked to use a service robot for recycling recommendations, where the robot is asked to recognize and sort an object.	Unpredictable responses.	[34]
Learning companion	Interview about human-like imperfections of intelligent learning companions that the participant already knows.	Any human-like imperfections.	[35]

Table 2.1: An overview of the different researches with the context, error, and scenario of the study.

The researches referenced in table 2.1 mostly conduct research on error situations with anthropomorphic robots. Therefore, we are not certain that these specific researches apply to non-anthropomorphic robots. To understand the impact of errors when using different levels of anthropomorphism in robots, more research into anthropomorphism in error situations needs to be done.

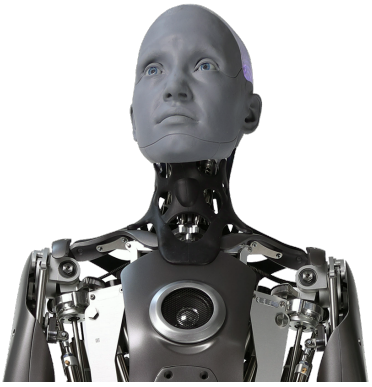
2.2 Anthropomorphism in HRI

To understand anthropomorphism it is essential to understand what it entails. Anthropomorphism is the inclination to think about inanimate objects, animals, and others as having human characteristics to put their actions into a certain perspective [36]. These human characteristics include cognitive or emotional states. By attributing these characteristics to for example robots, their behaviour in social environments is rationalized [36]. Anthropomorphism has become an important part of HRI since it has been seen to improve the interaction between robot and user [37]. Anthropomorphic design and social human-like characteristics can be used to try to increase people's acceptance and familiarity with a robot [38]. The phenomenon of anthropomorphism stipulates the importance of the design of the robot. Anthropomorphic design lies in many aspects of a robot. The robot's physical shape, social cues, human-like interaction, and facial expressions are all part of it [38]. Choi and Kim [39] specifically identify appearance, interaction, and how those two correlate as the important aspects of anthropomorphism in HRI.

Fong et al. [40] identify four categories of embodiment in a robot: *anthropomorphic*, *zoomorphic*, *caricatured*, and *functional*. See Figure 2.2. Where *zoomorphic* refers to an embodiment that resembles that of a creature or animal, *caricatured* points to a nonrealistic character embodiment, and *functional* applies to embodiment types that reflect the task that the robot is meant to perform. The form of a robot can help manage the expectations towards it, since the impression that the robot gives with its physical appearance impacts the interaction that follows [40]. Also, the level of anthropomorphism, meaning how closely it resembles a human, does play a role in the effects anthropomorphism can have. Mori [41] argued that when technology resembles humans too closely people have negative reactions towards the technology, like revulsion, since the small imperfections in the human-like design become more noticeable. Mori called this the "uncanny valley".

The consequences of anthropomorphic design in robots can be both positive and negative. For instance, the physical appearance of a robot can affect its perceived intelligence and intentions [46], [47]. This can be a positive effect when the perceived intelligence and intentions are in line with the actual capabilities of the robot, but when this is not the case this can have a negative consequence. Goetz et al. [48] stipulated that the design of anthropomorphic qualities in the robot should correspond with the capabilities the robot has. Furthermore, people can have different cultural, or individual preferences regarding the physical appearance of robots [49], [50]. This makes it even harder to optimally design a robot.

The positive effects that are attributed to anthropomorphism in robots include but are not limited to desire, likeability, pleasantness, competence, sociability, and



(a) Anthropomorphic [42]

(b) Caricatured [43]

(c) Functional [44]

(d) Zoomorphic [45]

Figure 2.2: The four categories of robot embodiments identified by Fong et al. [40]

trustworthiness. The familiarity of a robot's appearance positively affects the desire people feel towards the robot as well as the accessibility of the robot [40]. The likability and pleasantness of an interaction can also be positively impacted by anthropomorphism. Eyssel et al. [51] found that participants rated the interaction as more pleasant and the robot as more likable when they expressed emotions through nonverbal cues compared to when they reacted without these cues. Additionally, research in a health interview setting showed that an anthropomorphic robot was viewed as more dominant, trustworthy, sociable, responsive, competent, and respectful [52]. Furthermore, trust is also positively impacted by anthropomorphic design in autonomous cars [53] and virtual agents [54].

However, negative effects can also occur due to anthropomorphic design in robots. For example, it can cause an elevated feeling of embarrassment in medical examinations [55]. Additionally, in stressful situations, like search and rescue contexts, people found non-anthropomorphic robots more calming than anthropomorphic robots [56].

In conclusion, a robot's appearance and behavior can impact how people feel about these robots. For that reason, anthropomorphism is relevant in the field of

social robotics. And therefore, anthropomorphic design should be carefully considered. When considering the anthropomorphic design, context and interaction with the robot should be at the center of this consideration. Most importantly, when designing a social robot the anthropomorphic design must be appropriate for the capabilities of the robot and thus meet the expectations that it creates for the user [36], [48].

2.3 User experience

It has become apparent that many aspects of the user interaction with a robot are influenced by anthropomorphism. Similarly, how people perceive robots when they make errors is an important aspect of HRI. Designing robots to meet all expectations and make people accept the social robot is key in HRI. To find if this is the case the User Experience is often used as a measurement. More specifically, the user's trust in the robot, and the likability associated with the robot, are important for successful social robot interactions. To understand the importance of these two aspects of user experience, it is necessary to look closer into user experience itself.

User experience is a broad term that is used across multiple fields. Hartson et al. [57] define user experience (not specific to robots) as: "the totality of the effect or effects felt by a user as a result of interaction with, and the usage context of, a system, device, or product, including the influence of usability, usefulness, and emotional impact during interaction and savoring memory after interaction". A similar definition of user experience is described in ISO 9241-210 [58], which specifies that user experience is "a person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service". Additionally, Weiss, Bernhaupt and Tscheligi [59] identify embodiment, emotion, human-oriented perception, a feeling of security, and co-experience as indicators of user experience in Human-Robot Interaction. This implies that all the above-mentioned indicators can contribute to a positive user experience. A positive user experience is very important for social robots to add value to people's lives [3], and thus is important when creating meaningful interactions between a human and robot.

Bevan [60] specifies that *usability in use* is an important aspect of UX, where the usability of a product is captured within the UX. Bevan adopts the ISO 9241-11 definition of usability that includes effectiveness, efficiency, and satisfaction. The last component, satisfaction, is a subjective matter by which the user experience can be measured. Whereas the first two components, effectiveness, and efficiency, are objective measures. Measuring the perceived user experience can be difficult since it is very subjective and can depend on the history, skills, and personality of

the user [61], [62]. However, because UX is a subjective experience it is important to have subjective measures in addition to the objective measures. It is also important to note that how effective and efficient a product is can influence the user's satisfaction with that product, thus influencing the subjective measure. A more comprehensive overview of usability can be seen in Figure 2.3. Usability and user experience are closely related and criteria used for usability can be used to look at certain aspects of user experience [58]. Therefore satisfaction is deemed a strategic way to collect information about the user experience [63]. Bevan [60] specified that there are consequences that can be used to measure the UX, these are called: the *measurable consequences of UX*. These consequences are connected to the subjective components of user experience and can be used to measure the UX based on user input. There is a clear distinction between the user experience itself and its consequences. The user experience only lasts during the interaction. After the interaction is finished, the user experience is as well. However, the measurable consequences of UX continue to exist after the interaction is finished. These subjective measurable consequences include trust, pleasure, comfort, and likability [60]. Indicating that a high level of trust, pleasure, comfort, and likability points to a good user experience. These four consequences are not the only ways to measure user experience, but they can be used to find the positive or negative effects of certain interactions on the UX in HRI.

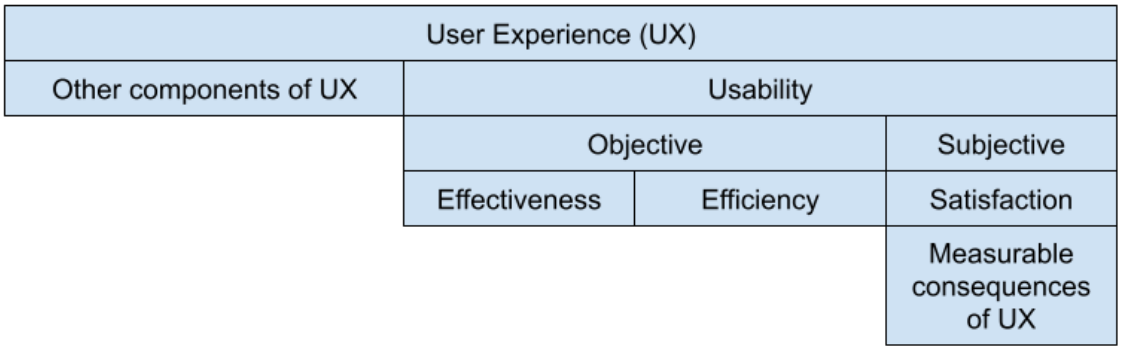


Figure 2.3: An overview of usability as a component of UX

The four measurable consequences of UX: trust, pleasure, comfort, and likability, are all important. However, their significance in a specific interaction depends on the context. To understand their significance an explanation of each consequences is given.

Pleasure

Pleasure is seen as an element that can influence the user's acceptance of a robot [64], [65]. Heerink et al. [64] point to the definition "feelings of joy/pleasure associated with the use of the system" for perceived enjoyment. Thus, in the context of social robot interaction, the terms "enjoyment" and "pleasure" are seen as the same feeling. Context is relevant to identify the importance of pleasure. For example, pleasure is less important when looking at the context of healthcare, here the functionality of the robot is more important. While in entertainment robots, pleasure is part of the functionality, making pleasure more important in this context.

Comfort

A good user experience can be characterized as an intuitive and comfortable one [66]. Comfort is specifically important in contexts in which a robotic device needs to be used by somebody every day or for long periods, for example, with robotic prostheses or robotic cars. The Robotic Social Attribute Scale (RoSAS) measures discomfort [67]. However, this is not the same as measuring comfort.

Moreover, comfort in HRI is a term that is closely connected to the movement of robots. Comfort is described as the satisfaction with *physical* comfort in an interaction [60]. Therefore, it seems logical that measuring comfort in HRI mostly pertains to physical attributes of the interaction, like movement. Consequently, much research in HRI about comfort concerns the movement of the robot. While research into interactions with little to no physical interaction does not take comfort into account when looking at the UX.

Trust

Lee and See [7] define trust in the context of automation as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". Hancock et al. [68] developed a triadic model of trust, where they found that human-robot trust was influenced by three components: Environmental factors, human-related factors, and robot-related factors. Notably, the context in which the interaction takes place gives information about the importance of trust in that context. For example, the risk level of the interaction can be an indicator of how important trust is in the interaction. For instance trust in automation is very important in the context of robot-assisted surgery, while a user's trust in domestic cleaning robots is less important for it to function correctly.

Different types of trust can be distinguished. For some types of trust, its level can change during an interaction, while other types are attributed to the characteristics

of robots and are thus stable over time [69]. An example of this would be the appearance of a robot. *Cognitive* and *affective* trust belong to the first category, where the level of trust can change during the interaction. Cognitive trust occurs when a person consciously chooses to trust somebody or something based on the information they have and it is dependent on the reliability and dependability of a specific partner. While affective trust is the trust that is placed on somebody or something based on their feelings and emotions towards that thing or person [69], [70]. Affective trust can also be influenced by information, similar to cognitive trust. The difference between the two lies in the reason for not trusting somebody or something, either due to conscious decision-making based on the available information or due to feelings and emotions.

Additionally, two domains of trust can be distinguished in human-robot interaction. Affective and cognitive trust can be divided into those two domains. The first domain applies to the social aspects of trust and is more focused on the relationship and morals. This regards the affective relationships and is related to affective trust. We will call this domain social trust. While the second domain applies more to the performance of the agent [71]–[73], which we will call performance trust. Cognitive trust can be allocated in this domain. For social robots, both of these domains should be taken into account [73]. And in error situations both of these domains are important.

Likability

Likability is seen as the positive first impression that is made [74]. A positive first impression can lead to positive future evaluations in the interaction between humans [5]. Furthermore, perceived likability has an influence on people's intention to use a robot [8]. Additionally, likability is often used as one of the measures for success in HRI [6]. In human-centered contexts, like entertainment or elderly care, likeability is an essential aspect. In contrast to contexts where people use certain robots to achieve a goal. Here likability may still be a design goal but is not necessary for the robot to do it's job. Examples of this are reception robots that show you the way or cleaning robots.

The measurable consequences of UX include four different aspects of UX. However, not all these aspects are relevant in the context of using anthropomorphic designs in HRI error situations. To understand what is and what is not important in the context of this research, the connections and dependencies of HRI error situations, anthropomorphism, and user experience are elucidated in the next section.

2.4 Effects on the perception of the interaction

If a robot makes an error, this influences the user experience of the interaction with that robot. The effect this has can depend on the type and severity of the error. For example, the user experience is negatively impacted by technical errors, while interaction or service failures have less of a negative impact [14]. Similarly, as discussed in section 2.2, the anthropomorphism of a robot can also influence the user experience. Errors and anthropomorphism are connected when talking about human-robot interaction, as both can influence the user's perception of the interaction. Not only do they both influence the user's perception of the interaction, but they can influence each other as well. Similarly, there are other components that are affected by errors and there are components that affect the user's perception of errors.

2.4.1 Components that affect the user's perception of errors

Multiple components can be identified that affect the user's perception of errors. First of all, the behavior of the robot can influence the way that errors are perceived. Forewarning and recovery strategies can reduce the negative consequences an error has on the user's perception [27]. For example, error recovery can lessen the negative impact an error has on the user's trust in the robot. This specifically holds when the consequences of the error are not very severe [13]. As noted previously, severity is very context-dependent, since an error in a risky situation can be more severe than the same error in a non-risk situation. Additionally, in error situations, some recovery strategies have different effects on the user's perception of the robot. For example, Lee et al. [27] found that the apology recovery strategy was the only strategy that they tested that made the robot appear more competent and likable.

A second category that can affect the user's perception of errors entails human-related factors. For example, the user's expectation for a robot can impact the user's view of the robot. More specifically, when a robot acts out of accordance with the user's expectation, this can be seen as an error [75].

As a third category, the robot's appearance is something that influences the user's perception of errors. An example of this is the embodiment, and thus anthropomorphism, of a robot. Kontogiorgos et al. [9] found that in case of failure the embodiment of smart speakers has a negative effect on the user's intention to interact with the device again after the failure. This negative effect is absent when the robot is embodied in a human-like way. An anthropomorphic embodiment also caused a higher rating regarding perceived social presence and intelligence. Additionally, in high severity failure situations a human-like embodiment was found to be distracting.

2.4.2 Attributes that are affected by errors

A lot of research has been carried out into the effect that errors can have on the user's perception of a robot. These include the following attributes: the perceived trustworthiness [9], [21], [22], [24], [25], [30], [31], [34], [76], anthropomorphism [21], [29], [32], [33], human-likeness [25], likability [9], [25], [32], [33], intelligence [21], [24], [29], [32], sincerity [17], competence [24], [25], [35], responsibility [18], [24], agency [18], predictability [24], dependability [24], safety [21], social presence [9], attitude [77], sympathy [77], companionship [35], and patience [29]. Additionally, some research has investigated the effects errors can have on the user's emotional state towards the robot. This includes research into user's curiosity [34], engagement [34], future contact intentions [9], [24], [33], willingness to use the robot [21], familiarity with the robot [17], acceptance [30], [31], satisfaction [31], and discomfort [24].

From all the attributes that are affected by errors, some overlap is seen with the measurable consequences of the user experience: trust, likability, and (dis-)comfort. However, pleasure is not found in the overview of attributes because no research was found on this. Additionally, these measurable consequences of UX also relate to anthropomorphism. Moreover, when considering these researches the context and categorizations of the errors determine the outcomes. For that reason, it is important to consider the categorizations and the contexts along with any anthropomorphic qualities.

Trust

Trust is an essential part of human-robot collaboration [78], while also being a measurable consequence of user experience. Additionally, as can be seen in Section 2.4.2 and Section 2.2 trust in a robot is also influenced by errors and the level of anthropomorphism. People largely base the trust they have in a machine on their perception of the competence of that machine. When a sign of incompetence is seen, this negatively influences the trust in that machine [79]. For social robots, mistakes have been shown to negatively influence the perceived reliability and trustworthiness of the robot in home companion robots [30]. This implies that, if a robot is making an error, this results in a decrease in the trust for that robot. However, trust is a complicated construct that is influenced by a lot of different components. One of these influences is the kind of trust with regards to the three categorizations of trust as discussed in chapter 2.1.1.

Furthermore, there is a correlation between error severity and the loss of trust in a robot. A higher loss of trust was found when the error was more severe [22], [80]. However, this does not hold for collaborative tasks, where both minor and se-

vere errors do not cause a significantly different level of trust compared to situations without errors [24], [25].

Additionally, Flook et al. [21] investigated the effects of different types of errors. Specifically, cognitive/decision level errors vs. technical failures. They found no difference in the level of trust between the two error types. However, they did find a difference between no error and either of the two error types, where the presence of an error caused a loss of trust. Correspondingly, other researchers also identified that failures/errors cause a loss of trust compared to no failures/errors [30], [31], [34].

Besides errors, anthropomorphism can also influence the trust one has in a robot. Since the design of a social robot affects its perceived trustworthiness [81]. Moreover, behavior and anthropomorphism are identified as important factors in predicting the trust people have in a social robot [76]. An anthropomorphic agent-interlocutor in a self-driving vehicle elevated the user experience and trust in autonomous cars [82]. Furthermore, human-human trust relationships are often less sensitive to errors than human-automation trust relationships. The reason for this is that people perceive automated systems as more credible than humans [76], [83], [84], and therefore do not expect errors. Thus, when a social robot is perceived as anthropomorphic people may be more lenient when they make errors. Consequently, when a robot shows anthropomorphic behavior like demonstrating emotions and awareness of an error, this positively influences the user experience and trust. Such that when a robot shows regret this causes the user's dissatisfaction to lower and sometimes even forgive the robot altogether [31].

Likability

Trust and likability are both related to the user experience [60], however, while trust is negatively impacted by errors, likability can be increased when robots make errors [32], [33]. Specifically, it was found that social norm violations (SNV) and technical failures (TF) in a robot, had a positive effect on the likability of the robot [32]. The Pratfall Effect is given as a possible reason for this phenomenon [32]. The Pratfall Effect states that people's attractiveness increases when they make small mistakes [85]. Furthermore, the difference between the two types (SNV and TF) was not measured. However, the positive effect of errors on likability was not found in the contexts of collaborative tasks in the workplace environment [25]. Where it was found that there was no difference in likability found between no error situation, mild errors, and severe errors.

Anthropomorphism also influences the perceived likability of a robot. As anthropomorphic qualities result in a higher rating of the likability of a robot [86]–[88]. Since

people have a higher opinion of things that they see are similar to them, and when people anthropomorphize a robot they become more similar to them [89]. However, there is a limit to this, when the robot is too anthropomorphic and appears nearly human, this will have a negative effect on the likability and people will find the robot unlikable [90].

Comfort

Little investigation has been done in HRI into the impact of errors on the comfort of people during an interaction. The research that did look into discomfort did not find a significant correlation between errors (no error, low severity, and high severity errors) and the level of discomfort the participant felt [24].

Similarly, not much information can be found on the effects of anthropomorphism on comfort. Akin to likability, anthropomorphic qualities do not have a positive effect on comfort when taken to the extreme. Mende et al. [90] explain that in situations where the robot looks extremely like a human, the user will feel discomfort. Moreover, May et al. [91] argue that, although comfort is not explicitly mentioned, some researchers point to anthropomorphism having a positive effect on comfort.

Pleasure

Similar to trust and likability, pleasure is also affected by anthropomorphism. More specifically, anthropomorphism has a positive effect on pleasure [82], [92] and enjoyment [93]. However, no research was found concerning the effect of errors on pleasure.

It has become clear that not all four measurable consequences of UX are equally applicable to the context of error situations with social anthropomorphic robots. Comfort is mostly regarded in HRI when talking about movements, as discussed in section 2.3, it seems that comfort is less important in stationary robots. Error situations occur in many settings, not only with moving robots. Combining that knowledge, with the fact that little information can be found on the influence of errors and anthropomorphism on comfort, it is clear that for this study, which concerns stationary robots, comfort can be excluded. Similarly, pleasure will also be excluded from this study. Although pleasure is relevant in some contexts, for the purpose of this research in the context of errors and anthropomorphism, it has no added value to be investigated. The reason for this is that pleasure is not found in any research on attributes that are affected by errors. Furthermore, the effect of anthropomorphism on pleasure is similar to that on trust and likability, which makes pleasure a replaceable

component in that regard. Therefore only trust and likability will be considered in the study.

2.5 Research objective

Research into how errors affect trust and likability can be found in abundance. However, research on the combination of anthropomorphism and errors in HRI and their effect on trust and likability is still missing.

As discussed previously, the likability of a robot increases when errors are made [32]. This is in line with the Pratfall Effect [32]. The Pratfall Effect is specifically about very competent humans, who are liked better when they make small mistakes [85]. Notably, the Pratfall Effect specifically mentions that it regards people, not robots, and that the errors are small. Indicating that anthropomorphism and severity possibly play a significant role in this phenomenon. First of all, since the Pratfall Effect concerns humans, this effect may not translate to social robots that are not anthropomorphic. Second of all, the severity of the error could also influence the applicability of the Pratfall Effect. Meaning that in severe error cases the Pratfall Effect does not hold.

Apart from likability, anthropomorphism has also been shown to help lower the negative effect of errors on a person's trust [76]. And the anthropomorphic design of a robot could help create a situation where the user is more forgiving of errors [82]. Making anthropomorphism an important part when looking at trust in HRI error situations.

For these reasons, this study focuses on likability and trust in error situations. More specifically, this study aims to find an answer to the questions *To what extent do the appearance of a robot (high-anthropomorphic vs. low-anthropomorphic) and the error severity (high-severity vs. low-severity) influence the level of trust and likability of a robot in collaborative scenarios?* With this question, this study aims to get more insight into the influence of anthropomorphic design on the user's perception of the robot in error situations, and if this in turn is influenced by the severity of the error. Several hypotheses were formulated based on the literature that was reported in the related works section.

Anthropomorphism lowers the negative impact that errors have on trust [31], [76]. For collaborative tasks, the severity level of the error has no significant impact on the level of trust [24], [25]. For that reason, only the level of anthropomorphic design has an influence on the trust towards the robot in error situations. Therefore the first hypothesis reads:

H1: *In error situations, robots with a high level of anthropomorphic design will be trusted more than robots with a low level of anthropomorphic design, regardless of*

the severity level of the error.

Anthropomorphism is seen to help mitigate the loss of trust [31], [76], therefore it seems logical to presume that errors have a bigger negative impact on robots with a low level of anthropomorphic design than on robots with a high level of anthropomorphic design. Thus, the second hypothesis is:

H2: *In error situations, robots with a high level of anthropomorphic design will have a lower amount of trust loss than robots with a low level of anthropomorphic design.*

Anthropomorphic qualities result in a higher rating of the likability of a robot [86]–[88]. Therefore the third hypothesis is:

H3: *In error situations, robots with a high level of anthropomorphic design will be perceived as more likable than robots with a low level of anthropomorphic design, regardless of the severity level of the error.*

The Pratfall Effect accounts for people being liked better when they make small mistakes [85]. When robots with high anthropomorphic qualities make small mistakes it is hypothesized that this will have the same positive effect on the likability of robots as it has on humans. The pratfall effect does not apply to non-anthropomorphic robots since it only holds for humans. Additionally, mostly negative attitudes towards robots with a low level of anthropomorphism were witnessed in error situations [9]. Therefore the fourth hypothesis is:

H4: *In low severity error situations, robots with a high level of anthropomorphic design will have an increase in likability while robots with a low level of anthropomorphic design will have a decrease in likability.*

In the case of high severity error situations, the mistakes are not small and therefore it is hypothesized that the Pratfall Effect no longer holds in these situations. Thus the decrease in likability for high anthropomorphic and low anthropomorphic robots will be the same. For that reason the last hypothesis reads:

H5: *In high severity error situations, robots with a high level of anthropomorphic design will have a similar decrease of likability as robots with a low level of anthropomorphic design.*

Chapter 3

Method

This study aims to find an answer to the question: *To what extent do the appearance of a robot (high-anthropomorphic vs. low-anthropomorphic) and the error severity (high-severity vs. low-severity) influence the level of trust and likability towards a robot in collaborative scenarios?* To find an answer to this question a study was designed and executed where the comparison could be made between high- and low-anthropomorphic robots in high- and low-severity error situations.

3.1 Approach

A 2x2 study (high anthropomorphism vs. low anthropomorphism x mild errors vs. severe errors) was carried out using two different robots. The robots had different levels of anthropomorphism (high and low). Additionally, the robots executed an error. For the error, there were different levels of severity (high and low). During the study, the participant and the robot collaborated by playing a (virtual) game together. Each participant had to interact with both robots. During the game, an error would occur. The severity level of this error was the same for both robots. Thus, anthropomorphism was researched using a within-subjects setup and error severity was researched using a between-subjects setup. This setup was chosen because participants could form feelings towards the robots. This could in turn influence their opinion of the robot in the next interaction. This made it impractical to research anthropomorphism using between-subjects. Additionally, by letting the severity level stay the same for both robots, the results are easy to compare. The virtual game used a Wizard of Oz approach for the robot. Which meant that the researcher operated the robots without the knowledge of the participants. Before the game started the participants were asked to have a small conversation with both robots. In this conversation, the robot introduced themselves and asked if the participant was ready to play a game. After the participant answered the robot said that the

game would start soon.

3.1.1 The game

A virtual treasure hunting game was played by the participants. The goal of the game was to find the treasure and its corresponding three keys. The participant got to see five layouts of a room, where each room's layout differs slightly from the other rooms. In each of these layouts, three keys and one treasure chest were seen by the participant. However, in each of the layouts, the keys and treasure chest were located in different places. The participant was told that the robot only sees the layout of one room, but does not see the keys or the treasure in that layout. The robot and the participant needed to work together to find which of the layouts the participant sees matched the robot's layout. The participant could ask the robot questions about the layout that the robot sees to identify the correct layout. The robot could only answer with yes or no. After the first two questions, the robot asked the participant for the location of one key. The participant would then (orally) tell the robot their answer, which the robot would repeat. After every further question, the robot asked for another key location. The participant got to ask four questions in total. Only four differences could be found in the five layouts, which makes four questions enough to identify the right map. The sequence of the game can be seen in Figure 3.2. The participants thus needed to give the location of all three of the keys and the treasure. An example of this setup can be seen in figure 3.1. The actual pages given to the participants for the setup can be found in Appendix A in Section A.3. For each key that they correctly identify the location of, they get points, 100 points for the first key, 75 points for the second key, and 50 points for the last key. If they correctly identify the treasure chest they get 250 points. The game finishes after they guess the location of the treasure chest.

This game shows similarities with a map task often used in linguistic studies. However, this game does not have the same complexity in the answers. For that reason, the answers in this game are only yes/no, and thus the same for each participant. This aspect makes it easier to run a Wizard of Oz study since no complex answers need to be formed.

This game is considered a collaborative game between the robot and the participant. Collaboration between humans and robots requires a common goal for the human and robot to work towards, where it is most often the robot that assists the human with its tasks to achieve the goal [94]. In the case of this study, the common goal is winning the game by getting all of the points. The robot and the participant have to work together to achieve this goal since the participant needs the robot to answer their questions and fill in the right answers.

3.1.2 Manipulation

For this study, two different aspects were manipulated: the level of anthropomorphism and the error severity level.

For the anthropomorphic levels two different robots were used: a low-level robot, which was represented by a Smart Speaker (SS), the Google Home, and a high level, which was represented by a Furhat [95]. The SS and the Furhat both used the same voice to mitigate any differences caused by using different voices. Each participant played a game with both of these robots.

Additionally, two levels of severity were used for the errors. The high severity error caused the team to lose the game and miss out on 250 points, while the low severity error only cost them 50 points. These severity levels are in line with Laprie's idea of benign and malign failures [12]. Both errors consist of the robot incorrectly repeating the participant's answer and thus filling in the wrong answer. This amounted to the robot saying "I will fill this in for the location of the key: [wrong location]" for the benign errors, and the robot saying: "I will fill this in for the location of the treasure chest: [wrong location]" Both errors are of the same type, where the robot incorrectly identifies what the participant has said. The participants were made aware that their scores would be noted on a scoreboard. A gift card was raffled between the participants that scored in the top 3. This prize gave the participants an incentive to win the game. This was done to imitate using a robot for a specific goal, as this influences how an error is received by the user.

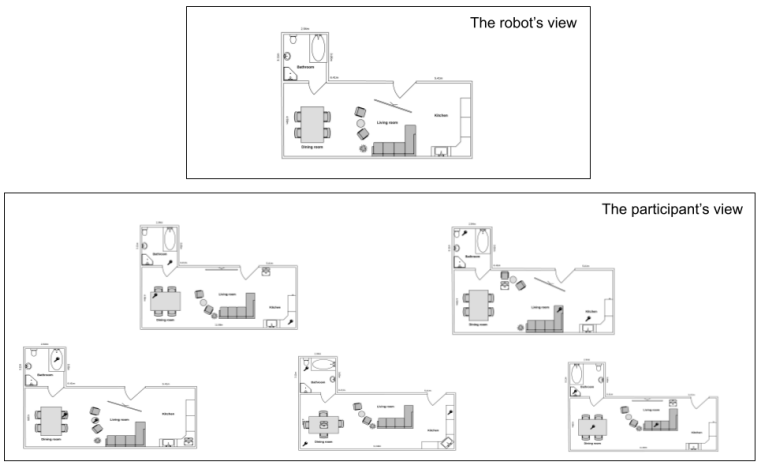


Figure 3.1: An example game overview

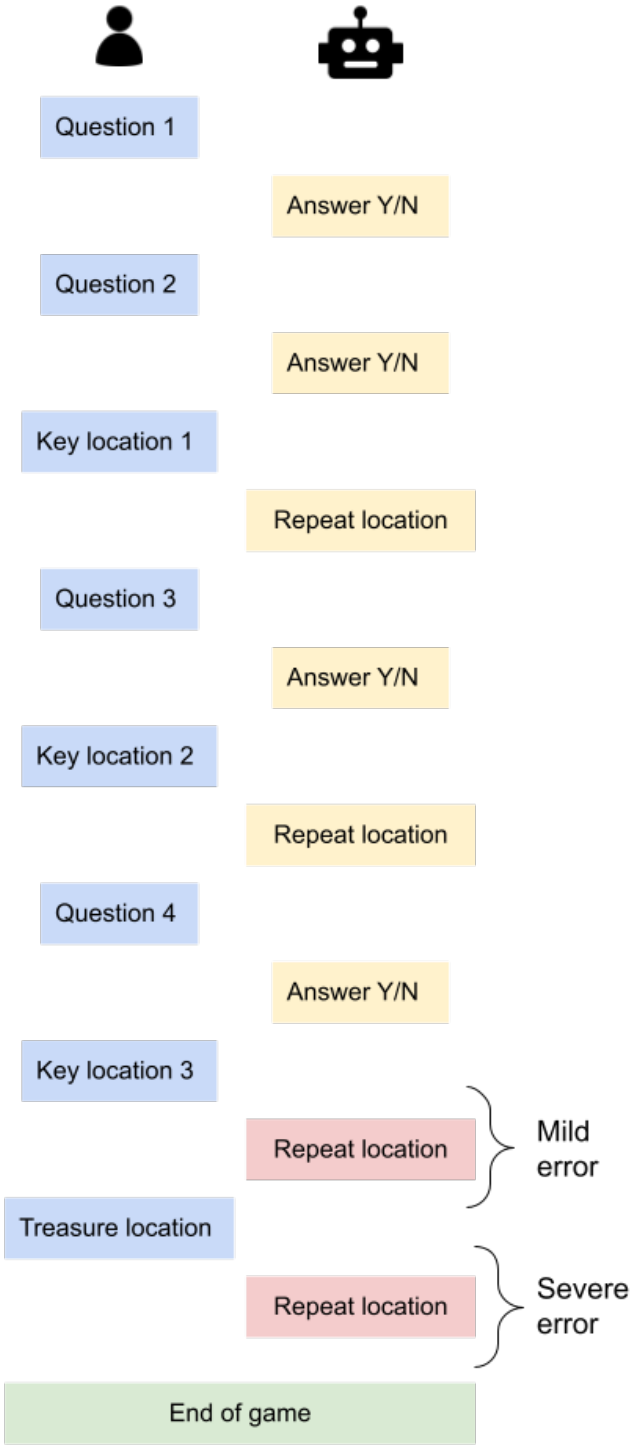


Figure 3.2: Overview of the sequence of the game. Where a robot would either execute a mild or a severe error by incorrectly repeating the location

3.2 Measurements

The two components that were measured were trust and likability. The measurement tools available to measure human-robot trust focus most on performance trust and less on social trust [73]. Schaefer [69] made a human-robot subjective trust scale of 40 items. Within these 40 items, they made a 14-item trust sub-scale focused on trust expectations. Of both these scales, the majority is focused on performance trust while only a few of the items focus on the social domain [73]. Malle and Ullman [73] propose a Multidimensional Measure of Trust where they look at both performance and social trust. They propose an 8-point rating scale with 20 items. Of these, 8 are based on performance trust and 12 are related to social trust. This Multidimensional Measure of Trust was used in this study as a subjective measure of trust. Similarly, likability was measured. A common and widely accepted way of measuring likability is using the Godspeed Questionnaire [74]. For that reason, it was used in this study. These two questionnaires were given to the participant four times, once before the game (which we will call the pre-study measurement) and once after the game (which we will call the post-study measurement) for each robot. Additionally, some extra questions regarding errors were asked in the questionnaire after the game. The participants were asked to rate the severity of the error they experienced. This was done to find outliers, give more context to the participants' answers and confirm the error severity level. The questionnaires can be found in Appendix A Sections A.1 and A.2.

Additionally, an interview was held with the participant after the experiment was over. This was done to gain insight into their feelings towards the robots and their motivation for these feelings. During this interview, the participant was asked, among other things, about their (previous) experience with the robots, if they were looking at the robot while talking, which robot they preferred and why, and they were asked about their feelings during the error.

3.3 Tools

For this study, the Furhat and the Google Home are used. The study has a Wizard of Oz setup where the researcher was using a program to make the Furhat and the Google Home talk to the participants. A program was written that could be executed on both devices. The program used the API available for the Furhat to contact the Furhat from a distance. The program incorporated the (US) voice named Matthew, which was available from the Furhat API. The program had a GUI with multiple buttons for the researcher to press, including a yes/no button, a button that repeated the (guessed) key location, and a button that repeated the (guessed) location of the

treasure. Additionally, the program had a button to start an introduction for both the Google Home and the Furhat in which they could introduce themselves and ask the participant if they were ready to play the game.

3.4 Procedure

Each participant was given a time slot of 1 hour in which they could complete the experiment. The Google Home and the Furhat were located next to each other during the experiment. When the participant entered, they were asked to sit in front of the two robots. The participants were then given an explanation of what was expected of them. It was explained that the study consisted of two parts. First, they would have a small conversation with each robot. After each conversation, they would fill in the first questionnaire *(containing the Godspeed questionnaire and the Multidimensional Measure of Trust)*. Second, the game would be explained and played with each robot. Afterward, the participants would fill in the second questionnaire (*containing the Godspeed questionnaire, the Multidimensional Measure of Trust, and additional questions as described in Section 3.2*).

After the explanation, the participants interacted with the first robot and filled in the first questionnaire (the pre-study measurement), which can be found in Appendix A Section A.1, and then did the same with the second robot. The order of the robots was randomized. The introductory interactions with both robots were held before any gameplay to keep the experience for the introductory interaction similar for both robots. This is done to minimize any changes in experience and emotion for the participant that can happen when they start playing the game.

After the introductory interaction, the participant got an explanation from the researcher about how the game worked. The game was then started and the participant played the game with the robot. The robot that was not used was covered up during the game. After the game was played the researcher gave the score of the first game and gave the participant the second questionnaire (the post-study measurement), which can be found in Appendix A Section A.2, After the participant filled out this questionnaire the second game started with the second robot. When the game ended the researcher gave the score of the second game and the total score. The participant was then asked to fill out the second questionnaire again but now for the other robot. After the questionnaire was filled out the researcher asked some additional interview-style questions. A visual overview of the procedure can be seen in Figure 3.3.

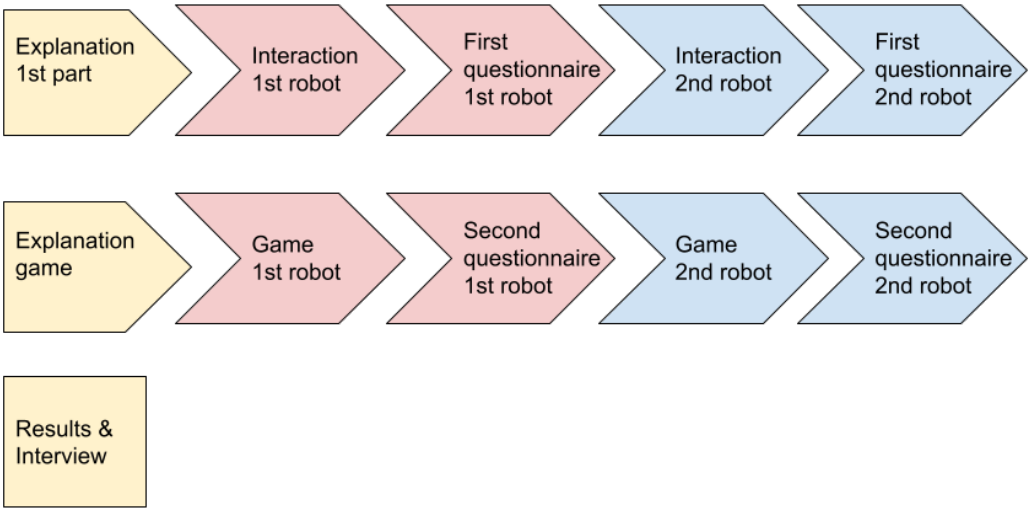


Figure 3.3: Overview of procedure, where the yellow blocks are done by the researcher, the red blocks represent actions for the participant with, and about the first robot and the blue blocks represent actions for the participant with and about the second robot. The order of the robots differed per participant.

3.5 Participants

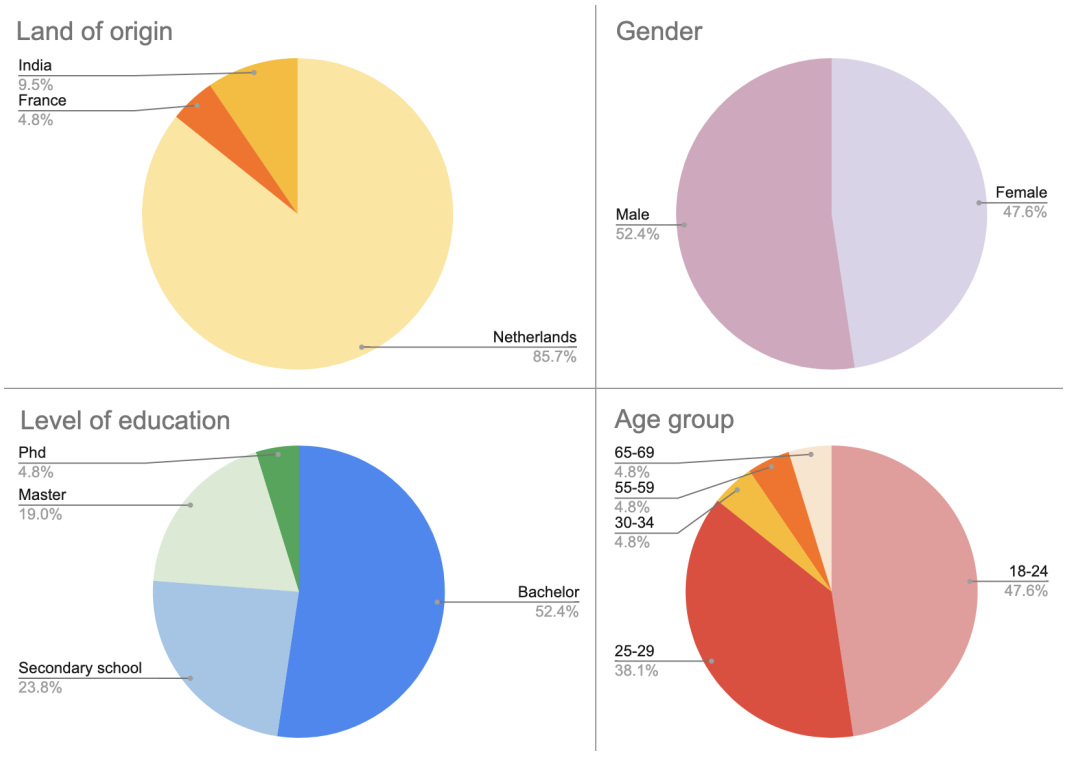


Figure 3.4: Demographics of the participant group, with n = 21

There are two different possible setups for a participant: the two robots with a high severity error in the interaction, or the two robots with a low severity error in the interaction. The goal was for each setup to be completed by 20 participants. Which would result in a total of 40 participants.

40 participants agreed to participate in the study. Of these 40 participants, 10 participants could not participate due to illness or other personal circumstances. Of the remaining 30 participants, 9 were excluded due to not noticing the errors. This became apparent during the post-measurement questionnaires, where they were asked if they had noticed an error and, if they did, which error they noticed. The post-measurements of these people were thus not influenced by any errors. These results could not be used in the study, because the likability and trust scores were not influenced by errors and thus not the same as the other results. This left 21 participants. Of these 21 participants, the average age of the participants was 28, and they all resided in the Netherlands at that time. The other demographic data can be seen in Figure 3.4.

3.6 Analysis

For the evaluation of the study, a mixed-method ANOVA was used to compare the questionnaire results for the likability and trust measurements, with an alpha of 0.05. Where the within-subjects factor is the level of anthropomorphism and the between-subjects factor is the error severity. This was done for the two dependent variables trust and likability. The results showed if any change in trust and/or likability is the result of the conjunction between anthropomorphism level and error severity level. Furthermore, the pre-study trust score and pre-study likability score were compared for the two robots by using a paired T-test. Additionally, the results from the interview were used to gain insight into the feelings of the participants towards the robots.

Chapter 4

Results

Before any analysis took place, a normality test was carried out on the data. All data was found to be normally distributed. The exact data normality tests can be found in Appendix B.

For the within-subject factor anthropomorphism (Furhat, Google Home), the differences between the pre-study measurement and the post-study measurement were taken as dependent variables for each of the levels of anthropomorphism. Data was gathered from a total of 21 participants. For the between-subject factor error severity (mild, severe) 11 participants experienced a mild error, and 10 participants experienced a severe error. The results of the user study can be sub-categorized into the two dependent variables: likability and trust.

4.1 Trust

4.1.1 Pre-study measurement

The trust participants felt towards the robots that was measured before the experiment began pointed to a preference for the Furhat. Where the Furhat scored an average of 4.3 for their trust score, while the Google Home had an average score of 3.8 as can be seen in Figure 4.1. Results of the paired t-test indicated that there is a significant small difference between the pre-study trust score for the Google Home (M = 3.8, SD = 1.1) and the pre-study trust score for the Furhat (M = 4.3, SD = 0.9), $t(20) = 2.2, p = 0.037$.

4.1.2 Post-study measurement

The post-study measurements of trust were analyzed using a mixed ANOVA test. The data that was found can be seen in Table 4.1 and can also be seen in Figure 4.2. From this data, we can see that the Furhat had a higher average trust rating after

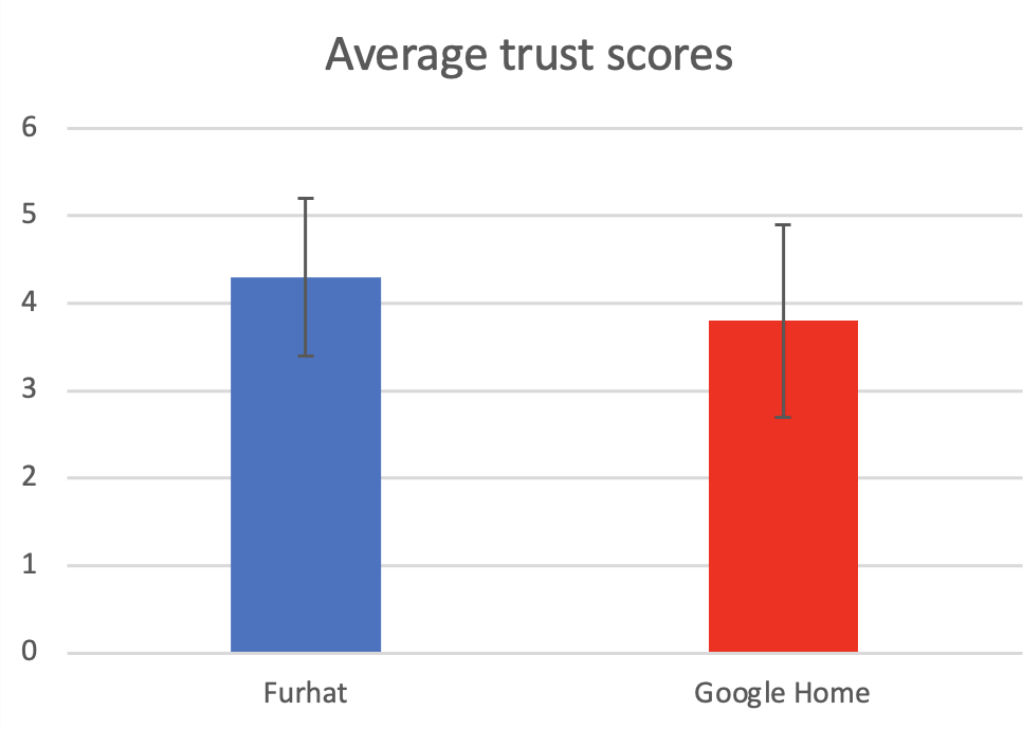


Figure 4.1: Average level of pre-study trust score per robot on a scale of 1 to 5, including the standard deviation as error bar.

errors (4.7) than the Google Home (3.5). Additionally, for the Furhat the average trust score was higher after mild errors compared to severe errors. However, for the Google Home the average trust score is higher after severe errors, although this difference is small.

For the ANOVA analysis, the following results were found. The *Tests of Within-Subjects Effects* analysis showed that there is a significant effect in the level of anthropomorphism on the trust scores overall: $F(1, 19) = 15.762, p < 0.001, \eta p^2 = 0.453$. However, there is no significant effect when looking at the interaction between anthropomorphism and error severity on the trust scores: $F(1, 19) = 2.086, p = 0.127, \eta p^2 = 0.118$. For the *Tests of Between-Subjects Effects* no significant effect was found of error severity on trust scores overall: $F(1, 19) = 0.323, p = 0.576, \eta p^2 = 0.017$. Additionally, the assumption that, for the between-participants variables, the groups that were compared have similar dispersion of scores is met according to Levene's Statistics with all p values being greater than 0.05.

4.1.3 Comparison pre-study vs. post-study

To compare the pre-study trust measurements with the post-study trust measurement, the average trust score of the pre- and post-study measurements was calculated for each participant. Meaning that the average was calculated from all the

Table 4.1: Descriptive Statistics of the average trust score post-study

	Error severity	Mean	Std. Deviation	N
Furhat (Average score)	Mild error	5.0208	1.23789	11
	Severe error	4.2750	1.26216	10
	Total	4.6657	1.27623	21
Google Home (Average score)	Mild error	3.4659	1.61544	11
	Severe error	3.6125	1.29428	10
	Total	3.5357	1.43676	21

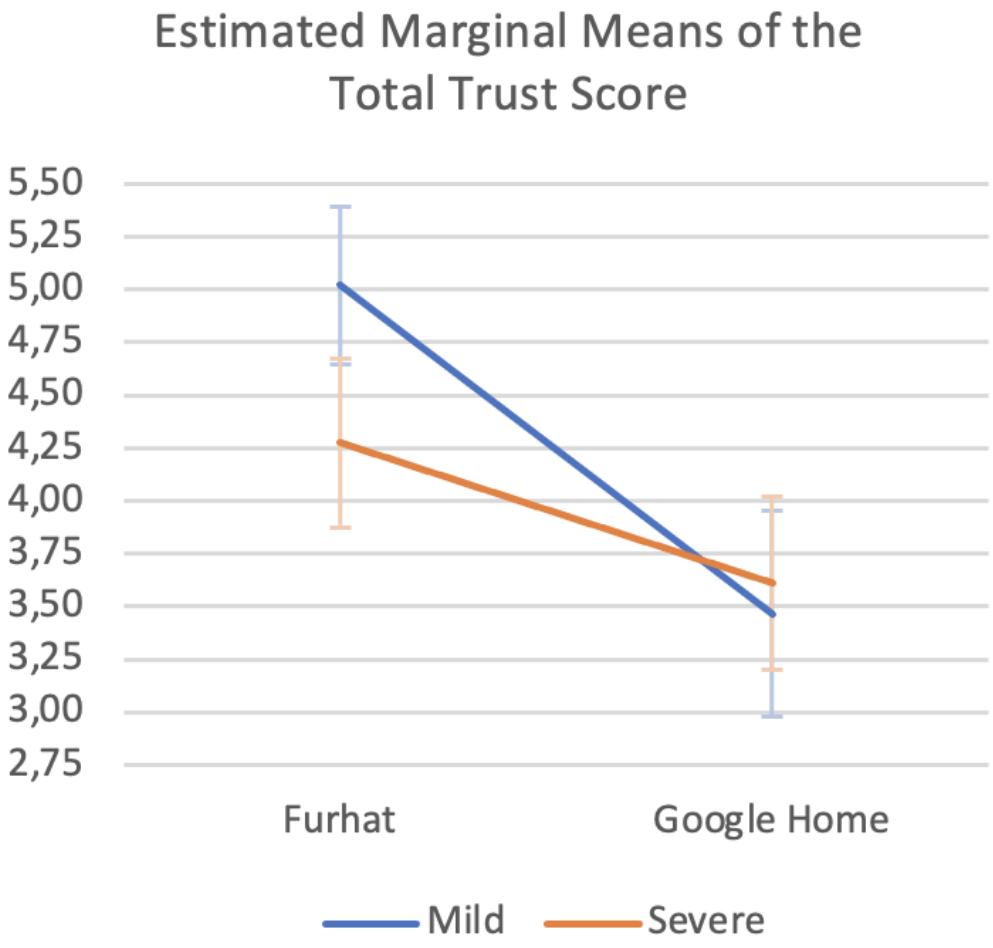


Figure 4.2: Estimated marginal means of the average total trust score, with the standard error of the mean (SEM)

trust scores one participant gave. These averages were then compared by subtracting the pre-study measurement from the post-study measurement to see the

average growth/loss. This made it possible to see what the difference was for one person between the average trust score they gave a robot pre-study and post-study. The descriptive statistics from this data can be seen in Table 4.2. The following observations can be made from this data: The average trust score for the Google Home decreased after the errors were made. This can be seen when looking at the average score, which is -0.2579. However, from the data, it can also be observed that after mild errors occurred the average seems to point towards a loss of trust: -0.5379, while the average difference in the trust after a severe error with the Google Home is positive: 0.05. For the Furhat the average trust level grew after an error occurred with 0.3244 points. Both mild and severe errors saw the trust grow on average, but mild errors caused a bigger average increase in trust (0.6023) than severe errors (0.0187). Overall the average score for trust growth towards the robots after the errors occurred is higher for the Furhat than it is for the Google Home. However, as can be seen in Figure 4.3, the average trust growth for both robots is similar after severe errors while after mild errors a big difference can be seen between the Google Home and the Furhat.

	Error severity	**Mean**	**Std. Deviation**	**N**
Google Home (post-pre)	Mild error	-.5379	1.01340	11
	Severe error	.0500	.92693	10
	Total	-.2579	.99531	21
Furhat (post-pre)	Mild error	.6023	1.20392	11
	Severe error	.0187	1.05740	10
	Total	.3244	1.14762	21

Table 4.2: Descriptive statistics table with the difference between the pre-study trust measurement and post-study trust measurement

The mixed ANOVA test that was performed on this data gave the following results. First of all, for the *Tests of Within-Subjects Effects* no significant effect was found in the level of anthropomorphism on the trust score growth: $F(1, 19) = 4.047, p = 0.059, \eta p^2 = 0.176$. Second, for the interaction between anthropomorphism and error severity in terms of trust score growth a significant difference was found: $F(1, 19) = 4.516, p = 0.047, \eta p^2 = 0.192$. Additionally, when looking at Levene's Statistics all p values were greater than 0.05, and thus the assumption that, for the between-participants variables, the groups that were compared have similar dispersion of scores is met. For the *Tests of Between-Subjects Effects* no significant effect was found of the error severity on growth of trust: $F(1, 19) = 0.000, p = 0.995, \eta p^2 = 0.000$.

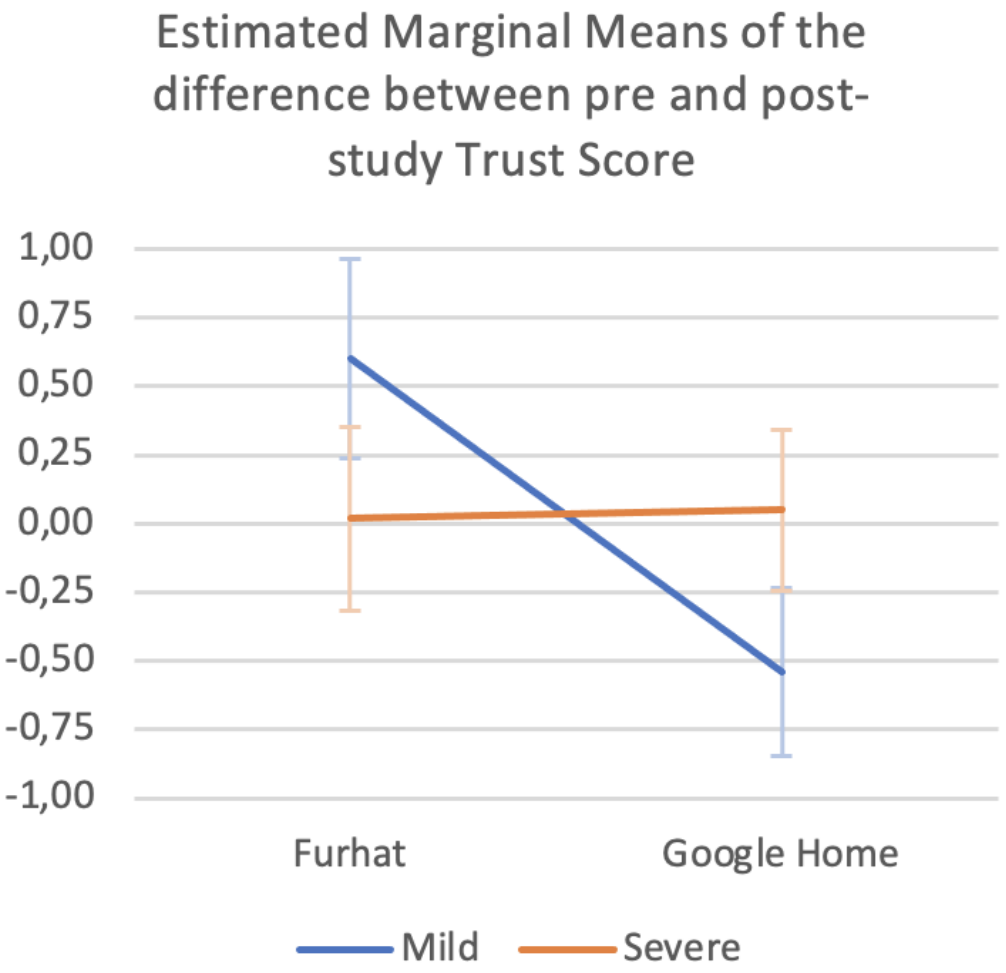


Figure 4.3: Estimated Marginal Means of the difference between the pre-study trust measurement and post-study trust measurement, with the standard error of the mean (SEM)

4.2 Likability

4.2.1 Pre-study measurement

For each robot, a pre-study measurement was done. The results from this measurement can give insights into the existing biases and preferences toward the two robots. For the likability of the robots, the pre-study measurement pointed to a preference for the Furhat. Where the Furhat's likability scored an average of 4.1 while the Google Home had an average score of 3.3 as can be seen in Figure 4.4. Results of the paired t-test indicated that there is a significant large difference between the pre-study likability score for the Google Home (M = 3.3, SD = 0.6) and the pre-study likability score for the Furhat (M = 4.1, SD = 0.6), $t(20) = 4.6, p < 0.001$.

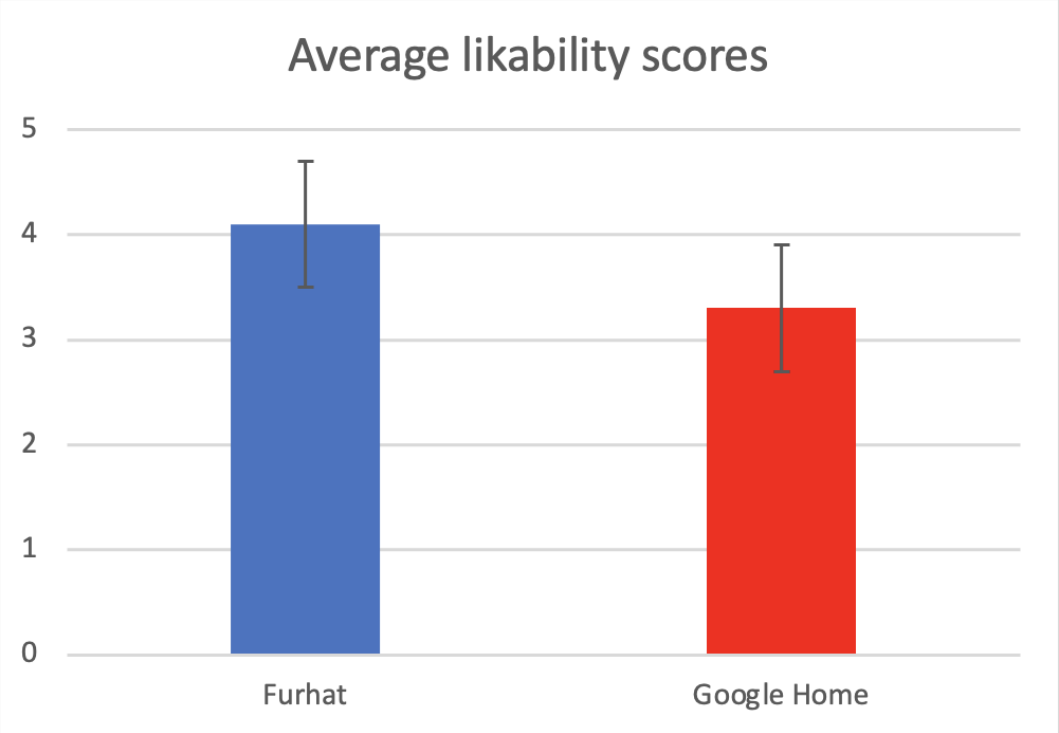


Figure 4.4: Average level of pre-study likability and score per robot on a scale of 1 to 5, including the standard deviation as error bar.

4.2.2 Post-study measurement

To analyze the post-study measurements of likability, a mixed ANOVA test was performed. The data used in this test is summarized in Table 4.3 and can also be seen in Figure 4.5. From this data, we can see that the Furhat had a higher average likability rating after errors (4.1) than the Google Home (3.4). Additionally, on average, the likability score was higher after mild errors compared to severe errors. However, for the Google Home this difference was very small.

For the ANOVA analysis, the following results were found. For the *Tests of Within-Subjects Effects* a significant effect was found in the level of anthropomorphism on the likability scores overall: $F(1, 19) = 10.071, p = 0.005, \eta p^2 = 0.346$. No significant effect was found when looking at the interaction between anthropomorphism and error severity: $F(1, 19) = 0.238, p = 0.631, \eta p^2 = 0.012$. For the *Tests of Between-Subjects Effects* no significant effect was found of error severity on likability scores overall: $F(1, 19) = 0.254, p = 0.620, \eta p^2 = 0.013$. Additionally, when looking at Levene's Statistics, it was seen that all p values were greater than 0.05, and thus the assumption that, for the between-participants variables, the groups that were compared have similar dispersion of scores is met.

Table 4.3: Descriptive Statistics of the average post-study likability score

	Error severity	Mean	Std. Deviation	N
Furhat (Average score)	Mild error	4.2182	.68384	11
	Severe error	3.9800	.84564	10
	Total	4.1048	.75530	21
Google Home (Average score)	Mild error	3.4000	.81486	11
	Severe error	3.3800	.76274	10
	Total	3.3905	.77065	21

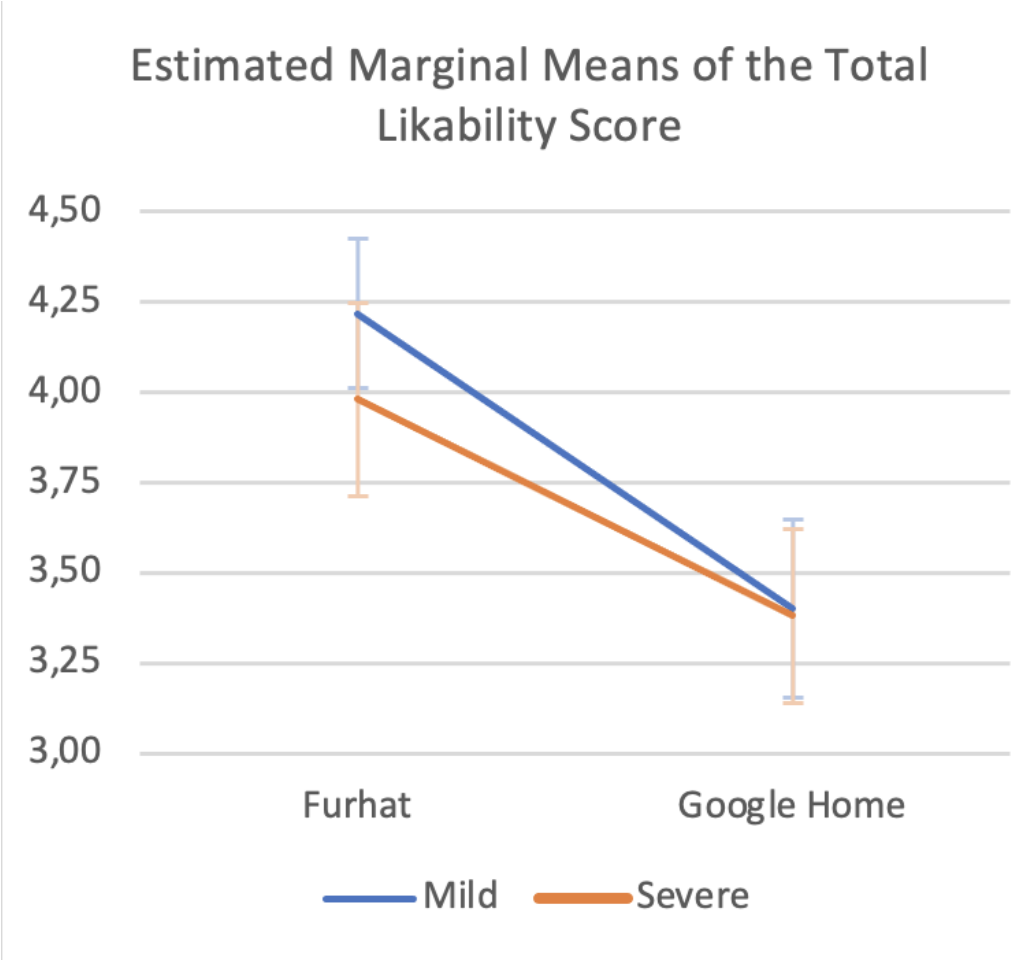


Figure 4.5: Estimated marginal means of the average total post-study likability score, with the standard error of the mean (SEM)

4.2.3 Comparison pre-study vs. post-study measurement

The comparison between the pre- and post-study likability measurements was done in a similar way to that of the trust scores. The descriptive statistics from this data

can be seen in Table 4.4. From this data, it can be seen that on average the likability score of both the Google Home and the Furhat grew after an error, with a total mean of 0.0952 and 0.0095 respectively. However, from this data, it seems that a mild error had a better effect on the likability score of the Google Home, with an average of 0.0364, than it had on the Furhat, with an average of -0.0182. The severe errors had a positive effect on the likability of both robots. Overall the likability score of the Google Home grew more than the likability score of the Furhat. This effect can also be seen in Figure 4.6.

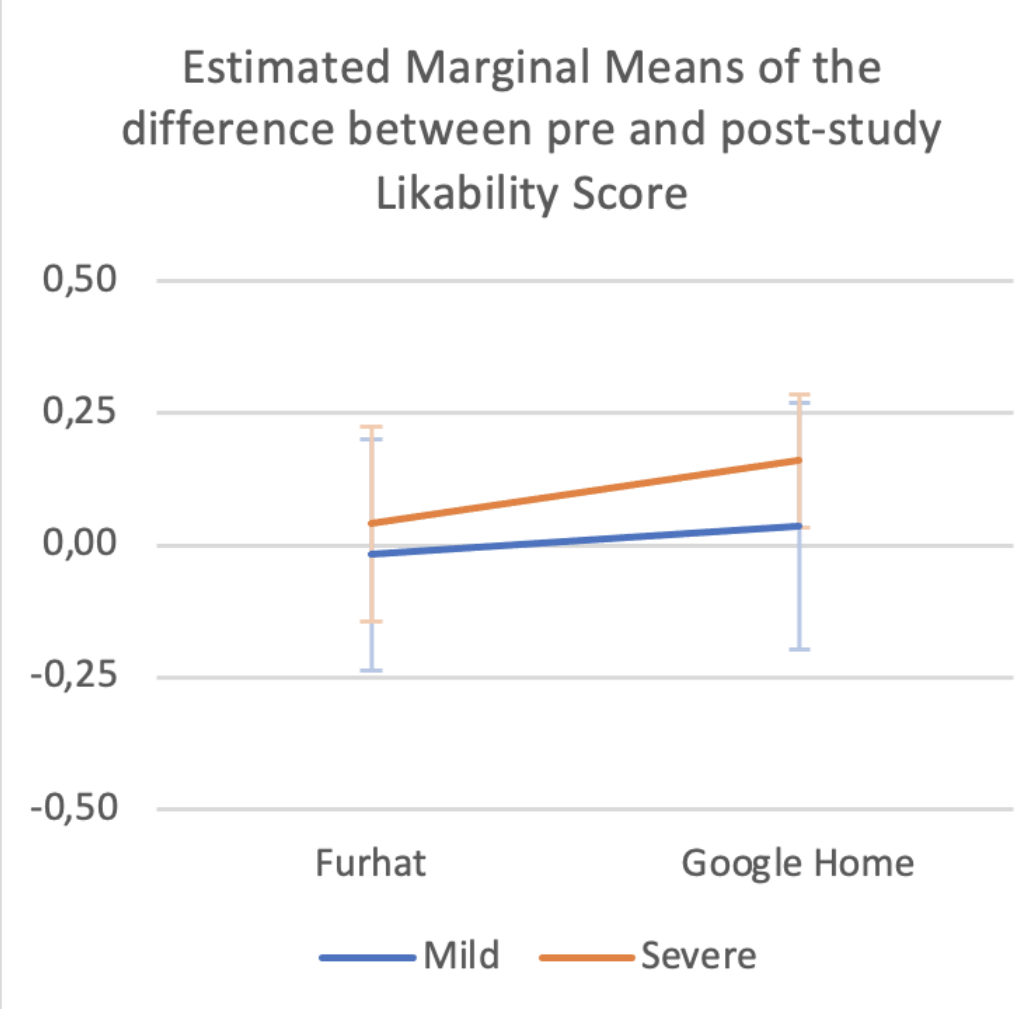


Figure 4.6: Estimated Marginal Means of the difference between the pre-study likability measurement and post-study likability measurement, with the standard error of the mean (SEM)

On the data, a mixed ANOVA test was performed. The assumption that, for the between-participants variables, the groups that were compared have similar dispersion of scores is met according to Levene's Statistics with all p values being greater than 0.05. The results of these ANOVA tests were that for the *Tests of Within-*

	Error severity	Mean	Std. Deviation	N
Google Home (post-pre)	Mild error	.0364	.77366	11
	Severe error	.1600	.39777	10
	Total	.0952	.61194	21
Furhat (post-pre)	Mild error	-.0182	.72363	11
	Severe error	.0400	.57966	10
	Total	.0095	.64335	21

Table 4.4: Descriptive statistics table with the difference between the pre-study likability measurement and post-study likability measurement

Subjects Effects there was no significant effect in level anthropomorphism in the growth/loss of likability scores: $F(1, 19) = 0.150, p = 0.703, \eta p^2 = 0.008$. Additionally, there was no significant interaction between Anthropomorphism and error severity in terms of growth/loss in likability scores: $F(1, 19) = 0.021, p = 0.886, \eta p^2 = 0.001$. For the *Tests of Between-Subjects Effects* no significant effect was found of error severity on growth/loss in likability scores: $F(1, 19) = 0.295, p = 0.593, \eta p^2 = 0.015$.

A complete overview of all results regarding trust and likability can be found in Figure 4.7.

4.3 Additional measurements

4.3.1 Severity

To check if the perceived severity level of the errors corresponds to the intended severity level, we can look at the following data. Each participant was asked how they would rate the severity of the error that they witnessed. The results of this question can be seen in Figure 4.8 and Figure 4.9. Here it is seen that the average severity rating of the mild errors is lower than that of the severe errors, as would be expected. However, when looking at Figure 4.9 it is seen that the median severity rating of both severity levels is the same at 5, on a scale of 1-5, meaning that there is a larger variety in severity rating of the mild errors than would be expected.

4.3.2 Observations

During the experiment several notable observations were made.

		Dependent Variables					
		Trust			**Likability**		
		Pre-study	Post-study	Comparison pre- and post- study	Pre-study	Post-study	Comparison pre- and post- study
Independent Variables	Level of anthropomorphism	A high level of anthropomorphism results in higher trust score	A high level of anthropomorphism results in higher trust score		A high level of anthropomorphism results in higher likability score	A high level of anthropomorphism results in higher likability score	
	Error severity						
	Interaction between anthropomorphism and error severity			For mild errors a *high level* of anthropomorphism results in trust growth, while a *low level* of anthropomorphism results in trust loss.			

Figure 4.7: Complete overview of all results for trust and likability. Where the green cells are significant results, the orange cells represent that no significant results were found and the grey cells mean that this was not researched.

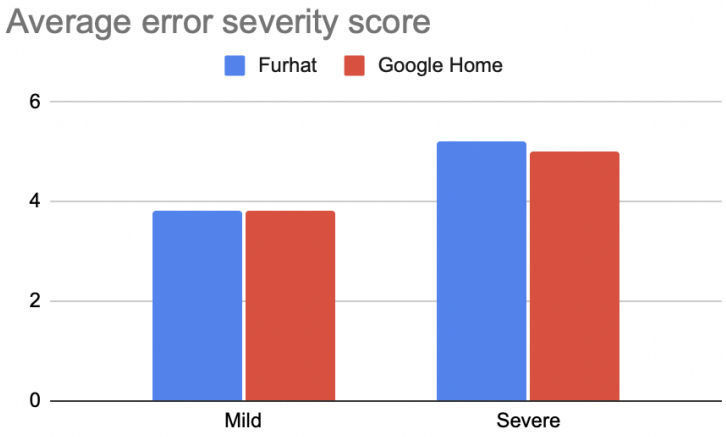


Figure 4.8: The average severity rating categorized by level of anthropomorphism

First of all, most participants were not looking at the robots, however, the participants reported looking more frequently at the Furhat than at the Google Home. Second, most (11) participants preferred the Furhat. Only 6 participants preferred the Google Home. The additional 4 participants did not have a preference. Third, from the excluded participants that did not notice any mistakes almost all said that they did not expect the robot to make mistakes. The reason they gave for this was that they felt that robots do not make mistakes. Additionally, multiple participants explained that they felt the Furhat was human-like when they made a mistake while they felt that the same mistake made the Google Home seem more robotic. Further-

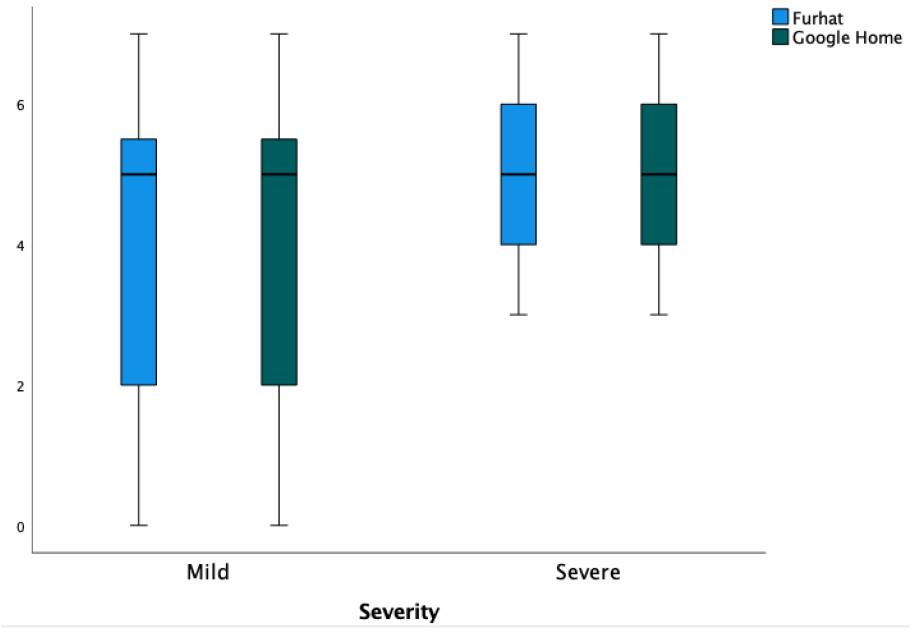


Figure 4.9: Box plot of the severity rating categorized by level of anthropomorphism

more, multiple participants gave the Furhat a very high score for likability during the pre-measurement questionnaire. This left them with little to no room to give a better score after the interaction had taken place. An example of this is that two participants gave the Furhat a perfect score for likability during both the pre- and post-study measurements. Meaning that this is represented in the data as neither an increase nor a decrease in likability. However, during the interview, they did explain that they liked the Furhat even more after the game.

Some of the reasons participants said they preferred the Furhat to the Google Home were that the Furhat was:

- More interesting
- Nicer
- More human
- More fun because he has a face
- More personal
- Charismatic
- Sweet
- Nice to look at
- Nicer due to the non-verbal feedback

Additionally, people reported that they preferred the Furhat. They felt that he made a human mistake, the interaction with the Furhat was nicer than that with the Google Home, and they liked to look at “someone” while talking.

The people that preferred the Google Home to the Furhat reported that they did not like the Furhat due to it having a face and that they felt watched. Additionally, one person said they preferred the Google Home because they had a better score in the game where they collaborated with the Google Home.

Some additional feedback that was given included that one participant explained that they expected more of the Furhat as it looked more complicated than the Google Home. Furthermore, two participants explained that they “did not mind the mistake as much the second time as it had already happened the first time” - Participant 7. However, Participant 8 had the opposite reaction: “I found the second mistake more frustrating than the first as it had already happened once.”

Chapter 5

Conclusion and Discussion

5.1 Discussion

Five different hypotheses were formed regarding the research question *To what extent do the appearance of a robot (high-anthropomorphic vs. low-anthropomorphic) and the error severity (high-severity vs. low-severity) influence the level of trust and likability towards a robot in collaborative scenarios?*:
H1: *In error situations, robots with a high level of anthropomorphic design will be trusted more than robots with a low level of anthropomorphic design, regardless of the severity level of the error.*
H2: *In error situations, robots with a high level of anthropomorphic design will have a lower amount of trust loss than robots with a low level of anthropomorphic design.*
H3: *In error situations, robots with a high level of anthropomorphic design will be perceived as more likable than robots with a low level of anthropomorphic design, regardless of the severity level of the error.*
H4: *In low severity error situations, robots with a high level of anthropomorphic design will have an increase in likability while robots with a low level of anthropomorphic design will have a decrease in likability.*
H5: *In high severity error situations, robots with a high level of anthropomorphic design will have a similar decrease of likability as robots with a low level of anthropomorphic design.*

5.1.1 Trust

A significant effect was found in the level of anthropomorphism on the overall average trust scores, this suggests that a higher level of anthropomorphism has a positive effect on the trust score compared to a lower level of anthropomorphism. A similar effect was observed during the pre-study measurements, as shown in Section 4.1.1. There it was seen that there was a significant difference between the level

of trust per robot. Implicating that robots with a higher level of anthropomorphism are trusted more than robots with a lower level of anthropomorphism regardless of errors. This result is similar to that found by Natarajan and Gombolay [76], who found that anthropomorphism lowers the negative impact that errors have on trust. Furthermore, for the interaction between anthropomorphism and error severity, no significant effect was found when looking at the average scores. This is in line with other research [24], [25], where it is explained that for collaborative tasks, the severity level of the error has no significant impact on the level of trust. Consequently, these results support H1 which states that robots with a high level of anthropomorphic design will be trusted more than robots with a low level of anthropomorphic design, regardless of the severity level of the error. *Thus, we can conclude that H1 is accepted.*

However, in terms of growth in trust, a significant difference was found in the interaction between anthropomorphism and error severity. Where in cases of mild errors the high-level anthropomorphic robot saw an increase in trust, while the low-level anthropomorphic robot saw a decrease in trust. This finding is in line with previous research [31] that found that high levels of anthropomorphism help minimize the loss of trust. However, this did not hold for severe error situations. The results of this study found that for the severe errors the levels of gained trust looked similar for both levels of anthropomorphism. This result was not completely in line with the research on severity levels for collaborative tasks by van Waveren, Carter, and Leite [24] and research by Sarkar, Araiza-Illan, and Eder [25]. However, it should be mentioned that their research did not specifically mention trust loss or gain. *Our results indicate that H2, which stated that robots with a high level of anthropomorphic design will have a lower amount of trust loss than robots with a low level of anthropomorphic design in error situations, is rejected.* As this hypothesis does not hold in severe error situations. Furthermore, it was seen that for the high-level anthropomorphic robot the trust in mild error situations grew compared to before the error. No similar effect was found in any research. As a negative impact on trust was expected after an error [30], [79]. However, Hamacher et al. did point out that anthropomorphic robots could be completely forgiven for their error [31].

5.1.2 Likability

Similarly to existing research [86]–[88], it was found that overall higher levels of anthropomorphism result in a higher rating of the likability of a robot. This was substantiated by the finding of a significant effect of the level of anthropomorphism on the overall average likability scores. Where the high-level anthropomorphic robot received significantly higher likability scores than the low-level anthropomorphic robot.

This also corresponds with the findings of Kontogiorgos et al. [9] that witnessed mostly negative attitudes towards robots with a low level of anthropomorphism in error situations. *This result supports H3, therefore this hypothesis is accepted.*

When analyzing the difference in likability before and after the errors occurred no significant difference could be found. For low error severity situations, it was hypothesized that robots with a high level of anthropomorphic design will have an increase in likability while robots with a low level of anthropomorphic design will have a decrease in likability. As the Pratfall Effect explains that people are liked better when they make small mistakes [85], and Mirnig et al. [32] insinuate that the Pratfall Effect also holds for anthropomorphic robots. However, the findings of this study are not in line with this assertion. *Therefore H4 is rejected.* A similar result was found for high severity error situations, where no significant difference was seen in the increase or decrease in likability after the error occurred between the two levels of anthropomorphism. *This means that H5 is accepted.*

However, a point of discussion for H4 and H5 is that the questionnaire did not always leave room for the participant to improve their likability score for the robot, as mentioned in point five of Section 4.3.2. Some participants gave a very high or maximum likability score for the high-level anthropomorphic robot during the pre-study measurement. These participants had little to no room to point out if there had been an increase in likability because the questionnaire did not give them more room. During the interview after the interaction, some participants did point out that they felt an increase in likability towards the high-anthropomorphic robot. However, this is not represented in the quantitative data due to shortcomings in the questionnaire.

Furthermore, another point of interest is the perceived error severity. It is stated that the Pratfall Effect only holds for mild errors. However, as can be seen in section 4.3.1, for mild errors there is a large variety in severity scores. This can be an indication that the mild error was perceived as severe by the participants. Making the Pratfall Effect not applicable to them.

5.1.3 Qualitative findings

An interesting finding was that most participants did not look at the robots when they were playing the game. Some participants even said that they did not look at the robots at all during the game. They were all more focused on the paper in front of them with different room layouts. However, all non-visual aspects of the robots (voice, type of answers, and error types) were identical. This points to the preference being a more mental preference than having anything to do with actual visual input. When looking at the reasons people gave for their preference for the Furhat several reasons have nothing to do with its appearance. Being nicer, more

personal, charismatic, and sweet are all descriptions of its personality. However, the personality of the Furhat was in no way different from the personality of the Google Home. This result points to the Furhat being seen as anthropomorphic. Duffy [36] explained that anthropomorphism is the inclination to think about inanimate objects, like robots, as having human characteristics to put their actions into a certain perspective. However, this was not done to the Google Home which points to it indeed having a much lower anthropomorphic level. Similarly, the mistakes that the high-level anthropomorphic robot made were referred to as human-like while the mistake of the low-level anthropomorphic robot was seen as robotic by some participants.

Nevertheless, the human-likeness of the high-level anthropomorphic robot was not always seen as a positive thing. As some participants had negative things to say about the anthropomorphic qualities of the robot. One participant reported the feeling of being watched, while others did not like the face of the robot.

5.2 Limitations and recommendations

The results of the study give insight into the effect of the appearance of a robot and the severity of an error on the trust and likability towards a robot in collaborative scenarios. However, the study and methodology had several limitations. First of all, the participant group used in this study was relatively small and only existed out of people living in the Netherlands who were proficient in English. Subsequently, the participant group largely existed of people in the age range of 18-29. Culture and sociodemographic factors can influence people's attitude towards robots [96], [97], therefore the lack of cultural and sociodemographic variety in the user sample can skew the results. For that reason, more research with a higher variety of cultures and sociodemographics within the participant group is needed before the results can be generalized.

Second of all, the frustration and the perceived severity level of an error are very subjective. This was also seen in the results for the error severity level scores in this study. In this study, a between-subjects design for the severity level was not used. Therefore, it has become very difficult to accurately compare the impact of error severity on the same robot. As each person has a unique understanding of the severity level. However, this setup was not chosen because it would be very hard to use the same robot twice, as the first interaction can influence the results of the second interaction.

Furthermore, in this study, only one robot was used per anthropomorphic level, which could lead to results specific to these robots, and not generalizable to other robots in the same anthropomorphic category. Furthermore, even though all the software was designed to be the same for both robots, the hardware was different

which could have resulted in differences. An example of this could be the different speakers, which could cause differences in the voice of the robots. Additionally, the existing experience people had with these robots before the study could also influence the results. As prior experience with a robot has an influence on the users' trust and general attitude towards the robot [98]. By adding more robots in each category these influences on the results could be diminished. Therefore similar research with different robots would be helpful to give a more complete image of the effects anthropomorphic levels can have on trust and likability in case of errors.

Another limitation of this study was that the measure of likability did not leave any room for improvement if the participant gave a 'perfect score' during the pre-study measurement. This made it hard to interpret the results. Therefore, future studies should use additional measurements in these cases where the participants are asked about the growth or loss in likability.

Additionally, not all participants noticed that the robots made mistakes. A reason that was given for this by some participants was that they did not expect the robots to make mistakes. They felt that robots did not make mistakes. In some cases, the participants heard the error but thought that they had made the mistake and that it was not the robot's fault. All these participants also said that they had no experience with either of the robots. This could be a reason for this mindset but requires more research.

5.3 Conclusions

This study aimed to find an answer to the question:
To what extent do the appearance of a robot (high-anthropomorphic vs. low-anthropomorphic) and the error severity (high-severity vs. low-severity) influence the level of trust and likability towards a robot in collaborative scenarios?

The results suggest that the appearance of a robot and error severity influence trust. Considering that the anthropomorphic qualities have a positive effect on the level of trust towards the robot, regardless of the error severity level. Additionally, for mild errors, anthropomorphic qualities also have a positive effect on trust growth. However, this does not hold for severe errors, where the trust growth is similar for both robots with both high and low levels of anthropomorphism. Concluding that the level of error severity does not play a role in if anthropomorphic qualities result in a higher level of trust overall. However, the level of error severity does play a role in if anthropomorphic qualities result in a positive *development* in trust after the error occurred, implicating that the trust growth/loss after an error depends on error severity.

Furthermore, the results suggest that appearance has an influence on the likability of a robot, but the error severity level does not. As the anthropomorphic qualities result in an overall higher likability score, regardless of the error severity. While for the *development* of likability scores no difference was seen between the two robots. Additionally, when comparing the effect of the error severity levels no difference was seen in the overall trust scores and the development of the trust scores. Therefore, it can be concluded that the level of error severity has no impact on the overall likability scores as on the development of the likability scores. Nevertheless, as can be seen in Section 5.1.2, there is some discussion on the correctness of this finding due to the limitations of the questionnaire. Furthermore, the appearance of the robot in terms of the level of anthropomorphism only affects the overall level of likability and not the development of the likability.

Bibliography

[1] J. Broekens, M. Heerink, H. Rosendal *et al.*, "Assistive social robots in elderly care: a review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.

[2] N. Savela, T. Turja, and A. Oksanen, "Social acceptance of robots in different occupational fields: a systematic literature review," *International Journal of Social Robotics*, vol. 10, no. 4, pp. 493–502, 2018.

[3] B. Alenljung, J. Lindblom, R. Andreasson, and T. Ziemke, "User experience in social human-robot interaction," in *Rapid automation: Concepts, methodologies, tools, and applications*. IGI Global, 2019, pp. 1468–1490.

[4] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "trust perception scale-hri"," in *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016, pp. 191–218.

[5] T. L. Robbins and A. S. DeNisi, "A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations." *Journal of Applied Psychology*, vol. 79, no. 3, p. 341, 1994.

[6] A. S. Arora, M. Fleming, A. Arora, V. Taras, and J. Xu, "Finding "h" in hri: Examining human personality traits, robotic anthropomorphism, and robot likeability in human-robot interaction," *International Journal of Intelligent Information Technologies (IJIIT)*, vol. 17, no. 1, pp. 19–38, 2021.

[7] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[8] D. Cameron, S. de Saille, E. C. Collins, J. M. Aitken, H. Cheung, A. Chua, E. J. Loh, and J. Law, "The effect of social-cognitive recovery strategies on likability, capability and trust in social robots," *Computers in Human Behavior*, vol. 114, p. 106561, 2021.

[9] D. Kontogiorgos, S. van Waveren, O. Wallberg, A. Pereira, I. Leite, and J. Gustafson, "Embodiment effects in interactions with failing robots," in *Pro-*

ceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–14.

[10] D. J. Brooks, "A human-centric approach to autonomous robot failures," Ph.D. dissertation, University of Massachusetts Lowell, 2017.

[11] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018.

[12] J.-C. Laprie, "Dependable computing and fault-tolerance," *Digest of Papers FTCS-15*, vol. 10, no. 2, p. 124, 1985.

[13] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, "Exploring the impact of fault justification in human-robot trust," in *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, 2018, pp. 507–513.

[14] S. Honig, A. Bartal, Y. Parmet, and T. Oron-Gilad, "Using online customer reviews to classify, predict, and learn about domestic robot failures," *arXiv preprint arXiv:2201.03287*, 2022.

[15] M. Giuliani, N. Mirnig, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations," *Frontiers in psychology*, vol. 6, p. 931, 2015.

[16] R. Ross, R. Collier, and G. M. O'Hare, "Demonstrating social error recovery with agentfactory," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3*, 2004, pp. 1424–1425.

[17] T. Gompei and H. Umemuro, "A robot's slip of the tongue: Effect of speech error on the familiarity of a humanoid robot," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2015, pp. 331–336.

[18] S. v. d. Woerdt and P. Haselager, "Lack of effort or lack of ability? robot failures and human perception of agency and responsibility," in *Benelux Conference on Artificial Intelligence*. Springer, 2016, pp. 155–168.

[19] D. Kontogiorgos, A. Pereira, B. Sahindal, S. van Waveren, and J. Gustafson, "Behavioural responses to robot conversational failures," in *2020 15th*

ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2020, pp. 53–62.

[20] L. Tian and S. Oviatt, "A taxonomy of social errors in human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 2, pp. 1–32, 2021.

[21] R. Flook, A. Shrinah, L. Wijnen, K. Eder, C. Melhuish, and S. Lemaignan, "On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based hri experiments trustworthy?" *Interaction Studies*, vol. 20, no. 3, pp. 455–486, 2019.

[22] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario," in *International Conference on Social Robotics*. Springer, 2017, pp. 42–52.

[23] M. Stiber and C.-M. Huang, "Not all errors are created equal: Exploring human responses to robot errors with varying severity," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 97–101.

[24] S. van Waveren, E. J. Carter, and I. Leite, "Take one for the team: The effects of error severity in collaborative tasks with social robots," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 151–158.

[25] S. Sarkar, D. Araiza-Illan, and K. Eder, "Effects of faults, experience, and personality on trust in a robot co-worker," *arXiv preprint arXiv:1703.02335*, 2017.

[26] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, ""i don't believe you": Investigating the effects of robot trust violation and repair," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 57–65.

[27] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 203–210.

[28] S. Engelhardt, E. Hansson, and I. Leite, "Better faulty than sorry: Investigating social recovery strategies to minimize the impact of failure in human-robot interaction." in *WCIHAI@ IVA*, 2017, pp. 19–27.

[29] O. Mubin and C. Bartneck, "Do as i say: Exploring human response to a predictable and unpredictable robot," in *Proceedings of the 2015 British HCI Conference*, 2015, pp. 110–116.

[30] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 1–8.

[31] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder, "Believing in bert: Using expressive communication to enhance trust and counteract operational error in physical human-robot interaction," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2016, pp. 493–500.

[32] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, "To err is robot: How humans assess and act toward an erroneous social robot," *Frontiers in Robotics and AI*, vol. 4, p. 21, 2017.

[33] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.

[34] E. Law, V. Cai, Q. F. Liu, S. Sasy, J. Goh, A. Blidaru, and D. Kulić, "A wizard-of-oz study of curiosity in human-robot interaction," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 607–614.

[35] H. Y. Kim, B. Kim, S. Jun, and J. Kim, "An imperfectly perfect robot: discovering interaction design strategy for learning companion," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 165–166.

[36] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 177–190, 2003.

[37] V. Groom, C. Nass, T. Chen, A. Nielsen, J. K. Scarborough, and E. Robles, "Evaluating the effects of behavioral realism in embodied agents," *International Journal of Human-Computer Studies*, vol. 67, no. 10, pp. 842–849, 2009.

[38] J. Fink, "Anthropomorphism and human likeness in the design of robots and human-robot interaction," in *International Conference on Social Robotics*. Springer, 2012, pp. 199–208.

[39] J.-g. Choi and M. Kim, “The usage and evaluation of anthropomorphic form in robot design,” 2009.

[40] T. Fong, I. Nourbakhsh, and K. Dautenhahn, “A survey of socially interactive robots,” *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[41] M. Mori, “Bukimi no tani [the uncanny valley],” *Energy*, vol. 7, pp. 33–35, 1970.

[42] E. Arts, “Ameca - engineered arts,” 2022. [Online]. Available: https://www.engineeredarts.co.uk/robot/ameca/

[43] Tinkerbots, “My first robot.” [Online]. Available: https://www.tinkerbots.de/produkt/my-first-robot/

[44] ABB, “Abb’s collaborative robot -yumi,” 2022. [Online]. Available: https://new.abb.com/products/robotics/collaborative-robots/irb-14000-yumi

[45] Aibo, “aibo,” 2022. [Online]. Available: https://us.aibo.com/

[46] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer, “Understanding social robots: A user study on anthropomorphism,” in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 574–579.

[47] M. L. Walters, K. L. Koay, D. S. Syrdal, K. Dautenhahn, and R. Te Boekhorst, “Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials,” *Procs of New Frontiers in Human-Robot Interaction*, 2009.

[48] J. Goetz, S. Kiesler, and A. Powers, “Matching robot appearance and behavior to tasks to improve human-robot cooperation,” in *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.* Ieee, 2003, pp. 55–60.

[49] V. Evers, H. Maldonado, T. Brodecki, and P. Hinds, “Relational vs. group self-construal: Untangling the role of national culture in hri,” in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2008, pp. 255–262.

[50] D. S. Syrdal, K. Dautenhahn, S. N. Woods, M. L. Walters, and K. L. Koay, “Looking good? appearance preferences and robot personality inferences at zero acquaintance.” in *AAAI Spring symposium: multidisciplinary collaboration for socially assistive robotics*, vol. 86, 2007.

[51] F. Eyssel, F. Hegel, G. Horstmann, and C. Wagner, "Anthropomorphic inferences from emotional nonverbal cues: A case study," in *19th international symposium in robot and human interactive communication*. IEEE, 2010, pp. 646–651.

[52] S. Kiesler, A. Powers, S. R. Fussell, and C. Torrey, "Anthropomorphic interactions with a robot and robot–like agent," *Social Cognition*, vol. 26, no. 2, pp. 169–181, 2008.

[53] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of Experimental Social Psychology*, vol. 52, pp. 113–117, 2014.

[54] E. J. De Visser, S. S. Monfort, R. McKendrick, M. A. Smith, P. E. McKnight, F. Krueger, and R. Parasuraman, "Almost human: Anthropomorphism increases trust resilience in cognitive agents." *Journal of Experimental Psychology: Applied*, vol. 22, no. 3, p. 331, 2016.

[55] C. Bartneck, T. Bleeker, J. Bun, P. Fens, and L. Riet, "The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots," *Paladyn*, vol. 1, no. 2, pp. 109–115, 2010.

[56] C. L. Bethel, K. Salomon, and R. R. Murphy, "Preliminary results: Humans find emotive non-anthropomorphic robots more calming," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009, pp. 291–292.

[57] R. Hartson and P. S. Pyla, *The UX book: Agile UX design for a quality user experience*. Morgan Kaufmann, 2018.

[58] ISO DIS 9241-210, *Ergonomics of human-system interaction: Part 210: Human-centred design for Interactive Systems*. International Organization for Standardization, 2010.

[59] A. Weiss, R. Bernhaupt, and M. Tscheligi, "The usus evaluation framework for user-centered hri," *New Frontiers in Human–Robot Interaction*, vol. 2, pp. 89–110, 2011.

[60] N. Bevan, "Classifying and selecting ux and usability measures," in *International Workshop on Meaningful Measures: Valid Useful User Experience Measurement*, vol. 11, 2008, pp. 13–18.

[61] M. Hassenzahl and N. Tractinsky, "User experience-a research agenda," *Behaviour & information technology*, vol. 25, no. 2, pp. 91–97, 2006.

[62] J. Lindblom, B. Alenljung, and E. Billing, “Evaluating the user experience of human–robot interaction,” in *Human-Robot Interaction*. Springer, 2020, pp. 231–256.

[63] S. Borsci, S. Federici, S. Bacci, M. Gnaldi, and F. Bartolucci, “Assessing user satisfaction in the era of user experience: Comparison of the sus, umux, and umux-lite as a function of product experience,” *International journal of human-computer interaction*, vol. 31, no. 8, pp. 484–495, 2015.

[64] M. Heerink, B. Kröse, V. Evers, and B. Wielinga, “Assessing acceptance of assistive social agent technology by older adults: the almere model,” *International journal of social robotics*, vol. 2, no. 4, pp. 361–375, 2010.

[65] M. M. De Graaf and S. B. Allouch, “Exploring influencing variables for the acceptance of social robots,” *Robotics and autonomous systems*, vol. 61, no. 12, pp. 1476–1486, 2013.

[66] D. Sprute, K. Tönnies, and M. König, “A study on different user interfaces for teaching virtual borders to mobile robots,” *International Journal of Social Robotics*, vol. 11, no. 3, pp. 373–388, 2019.

[67] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, “The robotic social attributes scale (rosas) development and validation,” in *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, 2017, pp. 254–262.

[68] P. A. Hancock, D. R. Billings, and K. E. Schaefer, “Can you trust your robot?” *Ergonomics in Design*, vol. 19, no. 3, pp. 24–29, 2011.

[69] K. Schaefer, “The perception and measurement of human-robot trust,” 2013.

[70] M. G. Washington, “Trust and project performance: The effects of cognitive-based and affective-based trust on client-project manager engagements,” 2013.

[71] R. Stower, N. Calvo-Barajas, G. Castellano, and A. Kappas, “A meta-analysis on children’s trust in social robots,” *International Journal of Social Robotics*, vol. 13, no. 8, pp. 1979–2001, 2021.

[72] D. J. McAllister, “Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations,” *Academy of management journal*, vol. 38, no. 1, pp. 24–59, 1995.

[73] B. F. Malle and D. Ullman, “A multidimensional conception and measure of human-robot trust,” in *Trust in Human-Robot Interaction*. Elsevier, 2021, pp. 3–25.

[74] C. Bartneck, E. Croft, and D. Kulic, “Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots,” 2008.

[75] N. Mirnig, M. Giuliani, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, “Impact of robot actions on social signals and reaction times in hri error situations,” in *International Conference on Social Robotics*. Springer, 2015, pp. 461–471.

[76] M. Natarajan and M. Gombolay, “Effects of anthropomorphism and accountability on trust in human robot interaction,” in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 33–42.

[77] M. Ragni, A. Rudenko, B. Kuhnert, and K. O. Arras, “Errare humanum est: Erroneous robots in human-robot interaction,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 501–506.

[78] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, “Planning with trust for human-robot collaboration,” in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 307–315.

[79] B. M. Muir and N. Moray, “Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation,” *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[80] D. J. Brooks, M. Begum, and H. A. Yanco, “Analysis of reactions towards failures and recovery strategies for autonomous robots,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 487–492.

[81] S. Naneva, M. Sarda Gou, T. L. Webb, and T. J. Prescott, “A systematic review of attitudes, anxiety, acceptance, and trust towards social robots,” *International Journal of Social Robotics*, vol. 12, pp. 1179–1201, 2020.

[82] D. R. Large, K. Harrington, G. Burnett, J. Luton, P. Thomas, and P. Bennett, “To please in a pod: employing an anthropomorphic agent-interlocutor to enhance trust and user experience in an autonomous, self-driving vehicle,” in *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications*, 2019, pp. 49–59.

[83] D. A. Wiegmann, A. Rich, and H. Zhang, “Automated diagnostic aids: The effects of aid reliability on users’ trust and reliance,” *Theoretical Issues in Ergonomics Science*, vol. 2, no. 4, pp. 352–367, 2001.

[84] M. T. Dzindolet, L. G. Pierce, H. P. Beck, L. A. Dawe, and B. W. Anderson, "Predicting misuse and disuse of combat identification systems," *Military Psychology*, vol. 13, no. 3, pp. 147–164, 2001.

[85] E. Aronson, B. Willerman, and J. Floyd, "The effect of a pratfall on increasing interpersonal attractiveness," *Psychonomic Science*, vol. 4, no. 6, pp. 227–228, 1966.

[86] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot," in *International conference on social robotics*. Springer, 2011, pp. 31–41.

[87] A. Castro-Gonzalez, J. Alcocer-Luna, M. Malfaz, F. Alonso-Martin, and M. A. Salichs, "Evaluation of artificial mouths in social robots," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 4, pp. 369–379, 2018.

[88] S. J. Stroessner and J. Benitez, "The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features," *International Journal of Social Robotics*, vol. 11, no. 2, pp. 305–315, 2019.

[89] J. Van Doorn, M. Mende, S. M. Noble, J. Hulland, A. L. Ostrom, D. Grewal, and J. A. Petersen, "Domo arigato mr. roboto: Emergence of automated social presence in organizational frontlines and customers' service experiences," *Journal of service research*, vol. 20, no. 1, pp. 43–58, 2017.

[90] M. Mende, M. L. Scott, J. van Doorn, D. Grewal, and I. Shanks, "Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses," *Journal of Marketing Research*, vol. 56, no. 4, pp. 535–556, 2019.

[91] D. C. May, K. J. Holler, C. L. Bethel, L. Strawderman, D. W. Carruth, and J. M. Usher, "Survey of factors for the prediction of human comfort with a non-anthropomorphic robot in public spaces," *International Journal of Social Robotics*, vol. 9, no. 2, pp. 165–180, 2017.

[92] A. Kim, M. Cho, J. Ahn, and Y. Sung, "Effects of gender and relationship type on the response to artificial intelligence," *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 4, pp. 249–253, 2019.

[93] M. M. Van Pinxteren, R. W. Wetzels, J. Rüger, M. Pluymaekers, and M. Wetzels, "Trust in humanoid robots: implications for services marketing," *Journal of Services Marketing*, 2019.

[94] A. Bauer, D. Wollherr, and M. Buss, “Human–robot collaboration: a survey,” *International Journal of Humanoid Robotics*, vol. 5, no. 01, pp. 47–66, 2008.

[95] F. Robotics, “The world’s most advanced social robot,” Apr 2022. [Online]. Available: https://furhatrobotics.com/

[96] H. R. Lee and S. Šabanović, “Culturally variable preferences for robot design and use in south korea, turkey, and the united states,” in *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2014, pp. 17–24.

[97] P. Flandorfer, “Population ageing and socially assistive robots for elderly persons: the importance of sociodemographic factors for user acceptance,” *International Journal of Population Research*, vol. 2012, 2012.

[98] T. L. Sanders, K. MacArthur, W. Volante, G. Hancock, T. MacGillivray, W. Shugars, and P. Hancock, “Trust and prior experience in human-robot interaction,” in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 61, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2017, pp. 1809–1813.

Appendix A

A.1 Questionnaire before game

Please rate your impression of the robot on these scales:

	Not at all 0	1	2	3	4	5	6	Very 7
Reliable	o	o	o	o	o	o	o	o
Sincere	o	o	o	o	o	o	o	o
Capable	o	o	o	o	o	o	o	o
Ethical	o	o	o	o	o	o	o	o
Predictable	o	o	o	o	o	o	o	o
Genuine	o	o	o	o	o	o	o	o
Skilled	o	o	o	o	o	o	o	o
Respectable	o	o	o	o	o	o	o	o
Someone' you can count on	o	o	o	o	o	o	o	o
Candid	o	o	o	o	o	o	o	o
Competent	o	o	o	o	o	o	o	o
Principaled	o	o	o	o	o	o	o	o
Consistent	o	o	o	o	o	o	o	o
Authentic	o	o	o	o	o	o	o	o
Meticulous	o	o	o	o	o	o	o	o
Has integrity	o	o	o	o	o	o	o	o

Please rate your impression of the robot on these scales:

	1	2	3	4	5	
Dislike	o	o	o	o	o	Like
Unfriendly	o	o	o	o	o	Friendly
Unkind	o	o	o	o	o	Kind
Unpleasant	o	o	o	o	o	Pleasant
Awful	o	o	o	o	o	Nice

A.2 Questionnaire after game

Please rate your impression of the robot on these scales:

	Not at all 0	1	2	3	4	5	6	Very 7
Reliable	o	o	o	o	o	o	o	o
Sincere	o	o	o	o	o	o	o	o
Capable	o	o	o	o	o	o	o	o
Ethical	o	o	o	o	o	o	o	o
Predictable	o	o	o	o	o	o	o	o
Genuine	o	o	o	o	o	o	o	o
Skilled	o	o	o	o	o	o	o	o
Respectable	o	o	o	o	o	o	o	o
Someone' you can count on	o	o	o	o	o	o	o	o
Candid	o	o	o	o	o	o	o	o
Competent	o	o	o	o	o	o	o	o
Principaled	o	o	o	o	o	o	o	o
Consistent	o	o	o	o	o	o	o	o
Authentic	o	o	o	o	o	o	o	o
Meticulous	o	o	o	o	o	o	o	o
Has integrity	o	o	o	o	o	o	o	o

Please rate your impression of the robot on these scales:

	1	2	3	4	5	
Dislike	o	o	o	o	o	Like
Unfriendly	o	o	o	o	o	Friendly
Unkind	o	o	o	o	o	Kind
Unpleasant	o	o	o	o	o	Pleasant
Awful	o	o	o	o	o	Nice

Was the robot's performance flawless, or did you notice any mistakes?

_____

If you noticed mistakes, what mistakes did you notice?

_____

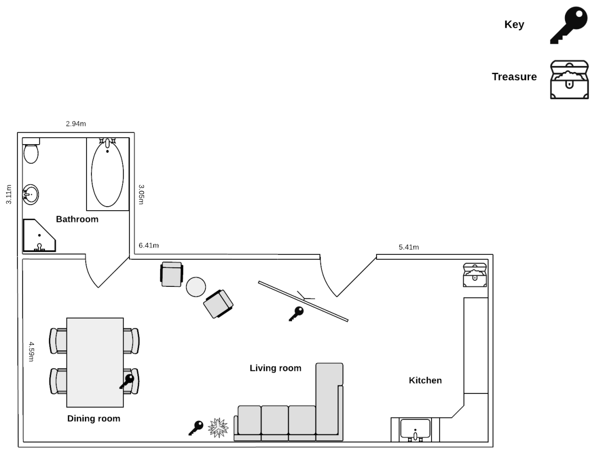
_____

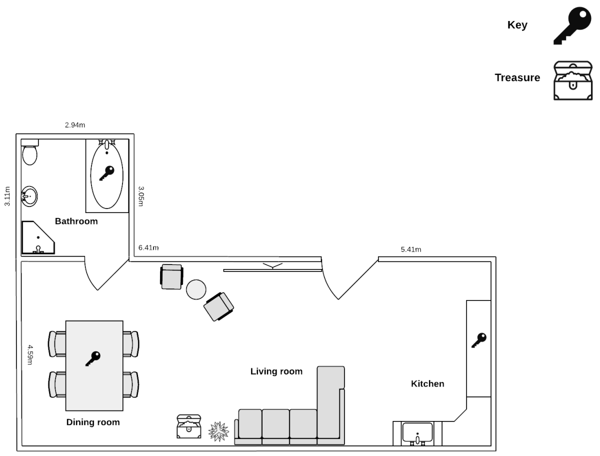
_____

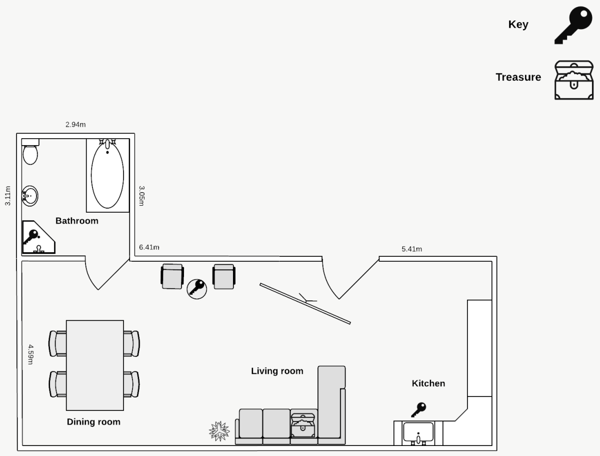
	Not at all 0	1	2	3	4	5	6	Very 7
How severe were/was the mistake(s)	o	o	o	o	o	o	o	o
How frustrating did you find the mistake(s)	o	o	o	o	o	o	o	o

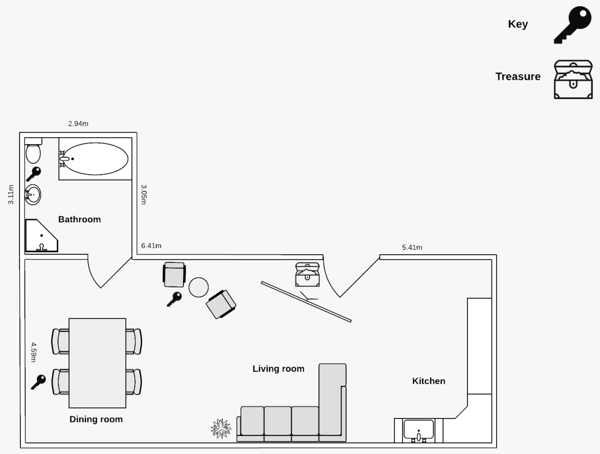
A.3 Treasure maps

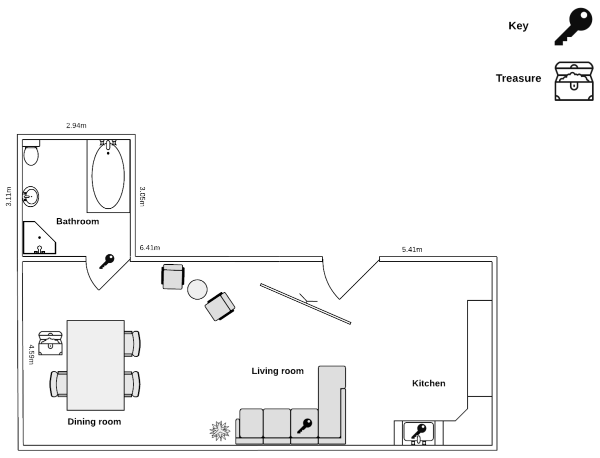
Two different overviews of treasure maps were used where each overview included maps of a unique room.







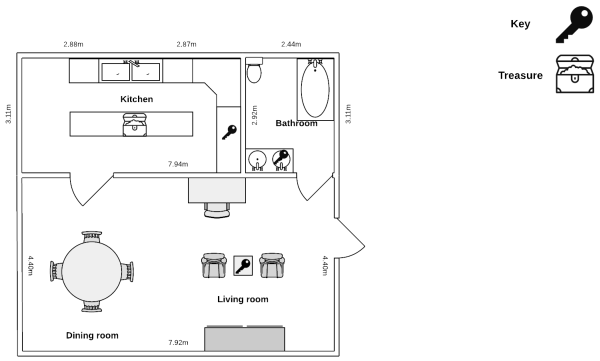


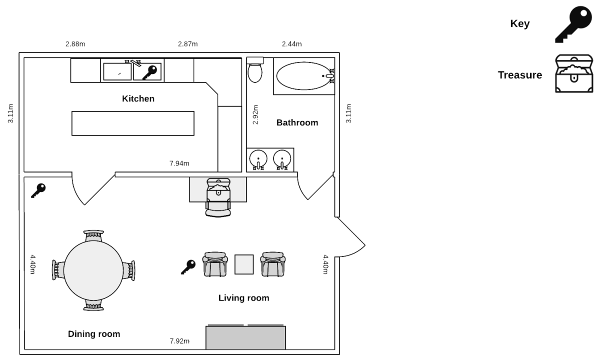


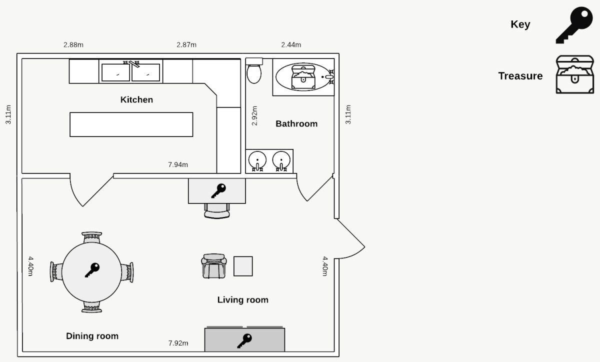
Rules:
- Only ask yes or no questions
- First key correct +100 points
- Second key correct + 75 points
- Third key correct +50 points
- For a correctly located treasure chest +250 points
- No questions may be asked about locations of keys or the treasure chest

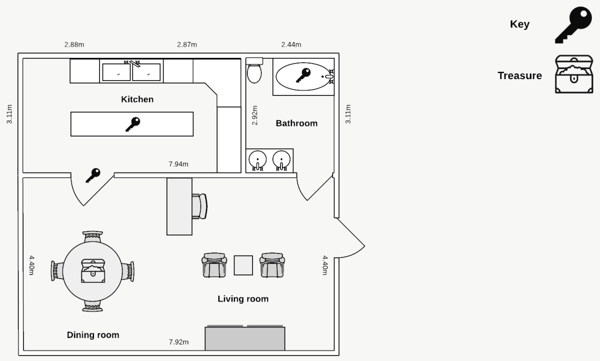
Goal:
- Try to find the location of the keys and the treasure chest of the map that the robot sees

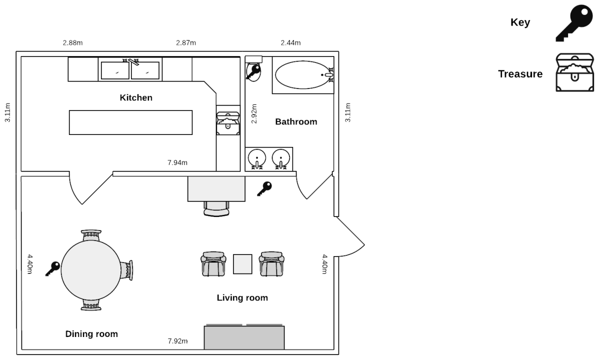
Game overview:
- ask 2 questions
- Give the location of the first key
- Ask 1 question
- Give the location of the second key
- Ask 1 question
- Give the location of the last key
- Give the location of the treasure chest
- END OF GAME











Rules:
- Only ask yes or no questions
- First key correct +100 points
- Second key correct + 75 points
- Third key correct +50 points
- For a correctly located treasure chest +250 points
- No questions may be asked about locations of keys or the treasure chest

Goal:
- Try to find the location of the keys and the treasure chest of the map that the robot sees

Game overview:
- ask 2 questions
- Give the location of the first key
- Ask 1 question
- Give the location of the second key
- Ask 1 question
- Give the location of the last key
- Give the location of the treasure chest
- END OF GAME

Appendix B

B.1 Normality assumptions

All gathered data regarding likability and trust was found to be normally distributed. As can be seen in Figure B.1 and B.2 all significance values from the Kolmogorov-Smirnov test and the Shapiro-Wilk test of normality are higher than the alpha of 0.05, meaning that we reject the null-hypothesis that the data does not have a normal distribution.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Furhat likability pre-test	.144	21	.200*	.953	21	.382
Furhat likability post-test	.128	21	.200*	.913	21	.064
Google Home likability pre-test	.139	21	.200*	.956	21	.436
Google Home likability post-test	.131	21	.200*	.952	21	.369

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure B.1: Tests of Normality for the likability data

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Furhat trust post-test	.108	21	.200*	.962	21	.565
Google Home trust pre-test	.120	21	.200*	.965	21	.615
Google Home trust post-test	.183	21	.065	.944	21	.261
Furhat trust pre-test	.111	21	.200*	.943	21	.253

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Figure B.2: Tests of Normality for the trust data