TU WIEN Informatics

Design and Evaluation of Non-Verbal Cues for the Robot Pepper

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Sarah Hanna Fischer, BSc
Matrikelnummer 01305079

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Mag. Dr. Margrit Gelautz
Mitwirkung: Darja Stoeva, MSc

Wien, 12. Oktober 2021

Sarah Hanna Fischer | Margrit Gelautz

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Design and Evaluation of Non-Verbal Cues for the Robot Pepper

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Media and Human-Centered Computing

by

Sarah Hanna Fischer, BSc
Registration Number 01305079

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Mag. Dr. Margrit Gelautz
Assistance: Darja Stoeva, MSc

Vienna, 12th October, 2021

Sarah Hanna Fischer	Margrit Gelautz

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Sarah Hanna Fischer, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 12. Oktober 2021

Sarah Hanna Fischer

Danksagung

Ich möchte mir sehr herzlich bei meiner Betreuerin Margrit Gelautz und Darja Stoeva für ihre Unterstützung bedanken. Sie waren immer für mich da und haben mich motiviert und mir bei Problemen weitergeholfen.

Acknowledgements

I want to warmly thank my supervisor Margrit Gelautz and Darja Stoeva for their support. You were always there for me and motivated me and helped me with any problems.

Kurzfassung

In den letzten 20 Jahren hat die Forschung zu nonverbalen Signalen für Roboter angefangen, animationsbasierte Techniken in den Designprozess einzubinden. Eines dieser Animationsprinzipien, welches erst wenige Forscher in diesem Kontext eingesetzt haben, ist *Übertreibung*. In dieser Arbeit sollen Fragen zu der Erstellung, Lesbarkeit und Wahrnehmung von übertrieben und nicht übertrieben dargestellten nonverbalen Signalen für den Roboter Pepper beantwortet werden. Die ausgewählten nonverbalen Signale Nicken, nach vorne Lehnen, Kopfschütteln und Gestiken wurden mit drei verschiedenen Methoden erstellt. Um die Forschungsfragen zu Lesbarkeit und Wahrnehmung zu beantworten, wurden zwei Online-Umfragen, welche Videos von dem Roboter beinhalten, und vier semi-strukturierte Interviews durchgeführt. Die erste Online-Umfrage konzentrierte sich auf Lesbarkeit von nonverbalen Signalen, welche von dem Roboter Pepper dargestellt wurden, und verglich verschiedene Methoden der Erstellung, sowie übertrieben mit nicht übertrieben dargestellten Signalen. Die zweite Online-Umfrage untersuchte, wie diese nonverbalen Signale beeinflussten, wie der Roboter von Menschen wahrgenommen wurde. Hierbei wurden zufällige, kontextbezogene und übertriebene Bewegungen mit Hilfe des Godspeed-Fragebogens verglichen. Die Interviews konzentrierten sich darauf, einen tiefergehenden Einblick in die Eindrücke von ausgewählten Umfrage-Teilnehmer*innen zu gewinnen, die die Videos, im Zuge derer der Roboter Pepper unterschiedliche nonverbale Signale darstellt, angeschaut hatten. Die zwei Godspeed-Variablen *Animacy* und *Perceived Safety*, welche mit *Belebtheit* und *Wahrgenommener Sicherheit* übersetzt werden können, waren statistisch signifikant. Sie zeigten eine höhere Bewertung von übertriebenen und kontextbezogenen nonverbalen Signalen gegenüber zufälligen nonverbalen Signalen. Diese Ergebnisse wurden durch aus den Interviews gewonnene quantitative Daten bestätigt.

Abstract

Over the last two decades, research on non-verbal cues for robots has started to include animation-based techniques in the design process. One of the animation principles that only few researchers have used in this context is *exaggeration*. This work seeks to answer questions about the creation, readability and perception of exaggerated and non-exaggerated non-verbal cues for the robot Pepper. The selected non-verbal cues nodding, leaning, head shaking and gestures were created using three different methods. To answer the research questions about readability and perception, two online surveys, each containing videos of the robot, and four semi-structured interviews were conducted. The first online survey focused on readability of non-verbal cues performed by Pepper and compared different methods of creation as well as exaggerated with non-exaggerated cues. The second online survey investigated how these non-verbal cues changed people's perception of the robot by comparing random, contextual and exaggerated movements, measured with the Godspeed questionnaire. The interviews centered around gaining more insight into people's experiences when watching the videos of the robot performing different non-verbal cues. The two Godspeed variables animacy and perceived safety were found to be statistically significant and showed higher scores for exaggerated cues as well as for contextual rather than for random cues. These results were supported by quantitative data from the interviews.

Contents

CHAPTER 1

Introduction

"The social robots market is expected [...] to reach a market size of US$ 912.488 million in 2026." [LLP21, Description para. 1]

The increasing growth of the social robots market can be felt as robots become more prevalent in areas such as teaching, coaching and healthcare, especially elderly care. They can be used as receptionists, entertainers, guides, or simply to keep someone company. As the robots in use in these areas increase in numbers and as the areas are becoming more diverse, the behaviour of robots needs to be designed to entail social skills, which will support the effectiveness of the interaction with humans in different contexts [SN19].

"[N]onverbal communication [...] is estimated to encompass more than 60% of all communicated meaning in human [interaction]." [SN19, page 575]

The researchers Mehrabian and Ferris [MF67], Mehrabian and Wiener [MW67], and Birdwhistell [Bir83] have tried to determine how much of human communication is non-verbal. While the exact percentage - such as the 60% mentioned in this quote - may vary, usually these researchers found evidence supporting the fact that a large part of human communication is non-verbal. This evidence is enough to warrant research into non-verbal cues for robots and suggests that it might be an even bigger part of interaction than verbal communication.

Therefore, interactions between humans and robots can become more effective by including an exchange of non-verbal cues. Gestures, for example, were used by Salem et al. [SER+13] to make a robot indicate where physical objects were by pointing at them. But non-verbal cues are not limited to being used to convey functional meaning. They can also be used to convey emotional meaning [SN19].

"Robot mood is contagious." [XBHN14, page 1]

The researchers Xu et al. [XBHN14] found that when they changed a robot's non-verbal cues in order to make it display a positive or negative mood, the self-reported mood of people changed in the same way. Craenen et al. [CDFV18], Andriella et al. [ASF+20], and Walters et al. [WSD+08] have tried to give robots different personalities, based on the Big-Five personality traits, by having them use different non-verbal cues, depending on their personality. They found that people generally preferred [CDFV18] or had better performance in a memory game [ASF+20] with robots with personalities similar to their own. Research like this shows that non-verbal cues can have an effect on how humans perceive robots. But more research in this area is needed, as one study by Li et al. [LJN15] shows that not all human traits are perceived equally in robots. In one study, Li et al. [LJN15] alternated interactions in which either the human or the robot displayed submissive or dominant behaviour and found that while people did not mind the dominant behaviour displayed by the human towards the robot, they did not appreciate the robot acting dominant towards the human.

1.1 Motivation

Robots are used more frequently and in a wider variety of fields than ever. These fields include computer science and engineering [TMS20], medicine [HGFT07], psychology [FWL+13], education or therapy [DW04], linguistics [FEN+04], coaching [CCM12] and many others [Dau03; GS07; SN19]. Especially, when they are used for teaching, coaching or as assistants, for example in elderly care, the robots are increasingly expected to have some sort of social intelligence. Dautenhahn [Dau03] even argues that for some applications social skills are vital to the effective use of the robot. One of these applications would be in elderly care, where the robot could be used as a companion for elderly people or as an assistant to the carers.

It has been shown by Terzioglu et al. [TMS20] and Dautenhahn [Dau07] that more socially capable robots are perceived more favourably by humans interacting with them and that social capabilities of robots can be improved with non-verbal cues. For example, Terzioglu et al. [TMS20] found that adding a fake breathing motion and eye gaze to the robot improved the human perception of the robot. This means that more human-like motions can improve how a robot is perceived. This raises the question whether all human movements are desirable in robots. Li et al. [LJN15] found that having the robot use body movements that make it appear dominant over a humans makes the robot receive worse ratings than when it is submissive. This shows us that not all human-like body movements are perceived as equally positive in robots and that we must carefully chose the ones we have our robots use. Therefore, more research is necessary to find out which human-like body movements are desirable in robots and which are not.

When looking at non-verbal cues, it is also important to consider the context in which they are expressed. Especially whether the non-verbal cues are contextual and therefore

fit the context or whether they are random. Bergmann et al. [BKE10] found that their baseline of random gestures, compared to their other methods of creating gestures, was rated worst in the categories of comprehension, vividness, likeability and competence. Babel et al. [BKM+21] found that the robot was given a higher anthropomorphism score when its non-verbal cue (gaze) was directed, rather than when it was random, during small talk with a human.

Animation principles have been used by researchers to try to improve the understandability and perception of non-verbal cues for robots [STH18]. For the process of finding studies which have used animation principles to implement in non-verbal cues for robots, the systematic review by Schulz et al. [STH18] can be considered particularly relevant. Those authors gathered studies that match this criterium and found only two publications which had looked at the principle of *exaggeration*, namely Gielniak and Thomaz [GT12] and Park et al. [PLC15].

Animation principles have often been used with success in the context of human-robot interaction (HRI). For this thesis, the animation principle of *exaggeration* was chosen, since we found that only few studies have looked into this specific principle in the context of social robots. Atkinson et al. [ADGY04] found that exaggeration increased the identification of emotional postures in videos performed by humans for all emotions except for sadness, which they found to be consistent with the idea that the feeling of sadness is generally associated with less movement [Dar15; Wal98]. This shows that readability of non-verbal movement can be increased with exaggeration in human-human interaction (HHI). Craenen et al. [CDFV18], while not calling it exaggeration, found that increasing amplitude or speed of an engaging gesture, resulted in higher scores in anthropomorphism, animacy and likeability of a robot.

1.2 Aim and Contributions

Overall, this research aims to answer the question: How does the animation-based principle of exaggeration change a human's perception of conversational non-verbal cues expressed by the robot Pepper? More specifically this is done by asking the following research questions (RQ):

RQ1: How do different ways of creating non-verbal cues for the robot Pepper impact their readability?

RQ2: How do random, contextual and exaggerated non-verbal cues influence a human's perception of anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of the robot Pepper?

RQ3: How do people describe their experiences of watching different non-verbal cues being expressed by the robot Pepper in the context of a conversation?

In order to answer these questions, we create readable non-verbal cues for the robot Pepper using the program Choregraphe [Roba]. The software is provided by the creator

of Pepper, Softbank Robotics [Robf], as a tool for creating motions and behaviours for the robot and contains a library of pre-implemented movements for Pepper. As mentioned in the previous section, it can have disadvantages to solely rely on the replication of human behaviours when creating non-verbal cues for robots [LJN15]. Therefore, the cues for this work are created in several different ways: 1) Using predefined cues from Choregraphe's library, that are designed for Pepper and therefore more machine-like; 2) Using replications of cues that are based on human-motion, by the authors of other papers; and 3) Cues that are more human-like and were created from-scratch using inspiration from the literature and human body motion. Additionally, in our work all the non-verbal cues undergo a process of exaggeration. The goal is to find out which cues are readable in the context of a conversation. This includes comparing three different ways of creating non-verbal cues and a comparison between the created cues and their exaggerated versions. The evaluation is done with an online survey to gather quantitative data and through statistical analysis.

Additionally, our research aims to find out how a person's perception of the robot Pepper is changed when it uses different non-verbal cues. To this end, another online survey and four semi-structured interviews are conducted, which allows for the collection of both quantitative and qualitative data.

We seek to gain insight into the creation, readability and perception of non-verbal cues for robots and into how animation principles, such as exaggeration, affect these three aspects. This will contribute to the knowledge around non-verbal cues for robots and animation principles being used in interactions between humans and robots.

The findings from the literature discussed in the previous section lead to the following hypotheses for RQ1 and RQ2:

Hypothesis 1: Exaggerated cues will be more readable than non-exaggerated cues [ADGY04].

Hypothesis 2.1: Anthropomorphism, animacy, likeability, perceived intelligence and perceived safety will be higher in the exaggerated cues than in the non-exaggerated cues [CDFV18].

Hypothesis 2.2: Anthropomorphism, animacy, likeability, perceived intelligence and perceived safety will be higher in the non-exaggerated than in the random cues [BKE10; BKM+21].

For RQ3 no hypothesis was formulated as it is a qualitative research question that will be investigated using data from the open-ended questions from the evaluation survey and four semi-structured interviews.

The perception of non-verbal cues for robots has been studied before [SER+13; Mav15; CDFV18; ASF+20; REF+20]. In order to provide new insights, this thesis will examine the effects of the animation principle of exaggeration applied in the creation of non-verbal cues for robots, as not many studies could be found on this topic. Exaggeration in the context of HRI has been studied by Gielniak and Thomaz [GT12] as well as Park

et al. [PLC15]. Park et al. [PLC15] focused on facial expressions of robots, which are not part of the research presented in this thesis, as the robot Pepper is not able to move any parts in its face to create different facial expressions. Additionally, Park et al. [PLC15] used exaggeration as a part of their non-verbal cues creation process, but did not draw comparisons to find out whether exaggerated non-verbal cues would be perceived differently. Gielniak and Thomaz [GT12] used the robot SIMON [Tec] and evaluated their own exaggerated and non-exaggerated cues. Contrary to our work, they did not compare different types of non-verbal cue creation and used a different questionnaire. They used memory testing to see if the exaggeration had any effect on memory, while this thesis focuses on user perception, with qualitative and quantitative data collection methods.

1.3 Structure of the Thesis

This thesis is structured as follows: Chapter 2 contains background information on HRI, including social robots. It also includes general information on non-verbal cues and animation techniques, followed by related work. Next, the methodology of the work is discussed in Chapter 3, which provides a detailed explanation of how non-verbal cues for the robot Pepper were created, how they were first validated using an online survey and then evaluated a second time via another online survey and semi-structured interviews. Chapter 4 reveals the results of our evaluation, including tables and charts to visualise the data. Chapter 5 is comprised of a discussion of the results and limitations of the work. The work concludes with a summary and suggestions for future work in Chapter 6. Appendix A contains copies of the online surveys and consent forms used in this work and Appendix B contains additional tables and charts.

CHAPTER 2

Background and Related Work

This chapter starts with an introduction to human-robot interaction and social robots. It offers definitions and examples as well as explanations. This is followed by an introduction into the concept of the uncanny valley, which is important when researching how humans perceive robots. A tool that is commonly used to evaluate the perception of robots is the Godspeed questionnaire: How it is constructed, how it works and how it has been used in the literature is explained in this chapter. Afterwards, research where videos of robots have been watched and rated by users is examined as well. This is followed by an explanation of non-verbal cues and how they have been adopted in human-robot interaction. Next, animation techniques and especially exaggeration are discussed: What they are, how they have been applied in hand-drawn animation and how they were adapted and are now sometimes used in the creation of non-verbal cues for robots.

2.1 Human-Robot Interaction

Although HRI is a research field which has not been around very long, the idea of the interaction between humans and robots has been on people's minds for decades. The idea of HRI and needing rules for it gained popularity in 1938, when famous science fiction author Isaak Asimov published a series of short stories which included the three laws of robot behaviour [Asi42; HGFT07].

These laws became so popular that there are multiple references to them in modern popular culture [HGFT07]. Examples are multiple references in the TV series *Dr Who* (1963) and *The Simpsons* (1989) as well as the movie *IRobot* (2004) and the video game *Portal 2* (2011). Asimov's [Asi42] laws are centered around the idea that robots who interact with humans need to abide by laws, so as to control them and make sure they do not harm the person they are interacting with. But the field of HRI encompasses much more than this. It is about finding out how one can make the interaction more

pleasant, effective and straight-forward, therefore increasing the willingness of people to interact with robots. HRI has been described by Bartneck et al. [BBE+20, page 7] as:

> "HRI focuses on developing robots that can interact with people in various everyday environments. This opens up technical challenges resulting from the dynamics and complexities of humans and the social environment. This also opens up design challenges—related to robotic appearance, behavior, and sensing capabilities—to inspire and guide interaction."

This description helps pinpoint the core ideas and challenges of HRI and shows how they emerged from different fields, such as design and robotics. The concept of design is present in the quote above where Bartneck et al. [BBE+20] explain what design challenges come up when working in HRI. The field of robotics is, though quite obviously related, harder to distinguish clearly from the field of HRI. A possible distinction was proposed by Bartneck et al. [BBE+20, page 6] in the following sentence:

> "One way to understand some key differences between the fields of HRI and robotics is that whereas robotics is concerned with the creation of physical robots and the ways in which these robots manipulate the physical world, HRI is concerned with the ways in which robots interact with people in the social world."

As Bartneck et al. [BBE+20] describe, the field of HRI is related to the field of robotics, but the difference lies within the focus of the research. HRI is also related to other fields, which include artificial intelligence and human-computer interaction (HCI). When focusing on the interaction, it seems only natural to take a look at robots whose main focus is social interacting with humans, social robots.

2.1.1 Social Robots

Social robots are a major part of HRI because social interaction with humans is at the core of their design and abilities. Social robots can be used to provide companionship or to work alongside humans, they can be teachers, coaches or provide help in other ways, such as giving directions [OKI+08]. Fong et al. [FND03] provide the following characteristics when describing what they call *socially interactive robots*, also known as social robots, in their work. Fong et al. [FND03, page 145] write: "Specifically, we describe robots that exhibit the following "human social" characteristics:

- express and/or perceive emotions;
- communicate with high-level dialogue;
- learn/recognize models of other agents;

- establish/maintain social relationships;
- use natural cues (gaze, gestures, etc.);
- exhibit distinctive personality and character;
- may learn/develop social competencies."

Social robots need to be more flexible than other robots, as they socially interact with humans. This interaction can include the robot acting as a partner or an assistant. The high flexibility is needed because humans are all very different and can interact in many different ways [FND03].

There are various kinds of social robots. Some examples are Wakamaru [Ind], NAO [Robb], Pepper [Robc], Asimo [Hon], Paro [PAR] and many more. The appearance of these robots can vary greatly, some having more of a humanoid body shape (e.g. Wakamaru, Nao, Pepper and Asimo), others resembling animals (e.g. Paro) and others being shaped like abstract forms (e.g. the half sphere shaped robot used by Zaga et al. [ZDL+17]).

In 1970, a Japanese researcher named Masahiro Mori published an article [MMK12] describing the uneasiness experienced by humans when they are confronted with a robot, or robot part, which looks very similar to a human or human body part but is actually synthetic. He calls this "the uncanny valley" and his graph visualising it can be seen in Figure 2.1. The uncanny valley did not interest many people at the time and the article did not receive a lot of attention. But many years later, the concept was rediscovered when people started to experience the phenomenon with modern technology, such as computer generated images in films, physical robots or virtual reality. This phenomenon is relevant to our research because it needs to be considered when working with robots. The creation of something that is extremely close to, but not actually human, can repel people and have the opposite effect we desire.

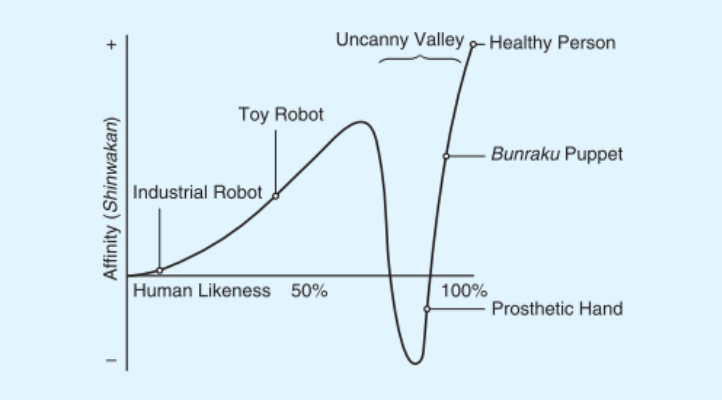


Figure 2.1: The uncanny valley as described by Mori et al. [MMK12].

2.1.2 Humans' Perception of Robots

Many different ways of measuring how humans perceive robots have been suggested by researchers over the years. But even if we reduce the amount of approaches and limit our scope to methods involving questionnaires, one important problem remains: many of the proposed questionnaires have not been evaluated regarding their validity and reliability. [BKCZ09]. Therefore, Bartneck et al. [BKCZ09] created and evaluated the Godspeed questionnaire in order to provide a way to measure humans' perception of robots with a questionnaire whose validity and reliability has been confirmed.

The Godspeed questionnaire uses five different measurements to rate how a robot is perceived: Anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Anthropomorphism describes as how human-like and animacy as how lively or lifelike the robot is perceived. These five measurements each use semantic differential scales with around 5 different words, depending on the measurement. A semantic differential scale is similar to a Likert scale but the difference is that the words are directly on the left and right of the numbers which can be chosen, see Figure 2.2. A Likert scale works by having a sentence and then always the same two words *Agree* and *Disagree*, between which the participant can choose a number. In both scales, the purpose is for the participant to chose a number between two words, depending on which word they think fits better. The Godspeed questionnaire and its official translations can be found on Bartneck et al.'s [BKCZ09] website[1].

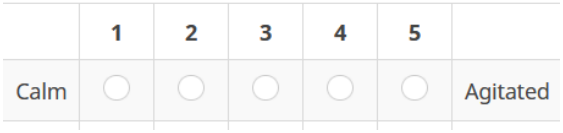


Figure 2.2: A semantic differential scale, where the participant chooses one of the numbers between the two words.

As mentioned before, the Godspeed questionnaire is often used when researchers want to answer questions about how participants perceive robots. However, it is especially useful when two or more different scenarios are compared, for example, being mirrored by a robot or not. To research this, the Godspeed questionnaire was used by Fuente et al. [FIPC15], who used the NAO robot to find out how being mirrored by the robot influenced the participant's perception of the robot during a conversation. They found that the robot's score for anthropomorphism, animacy and likeability increased when the robot used mirroring during the conversation. Zaga et al. [ZDL+17] had a different approach: They used the Godspeed questionnaire to find out if minimal gaze movements would impact a child's perception of the robot. They picked a low-anthropomorphic robot, which only consisted of a half-sphere shape with an eye, and found that even minimal social- and deictic-gaze already improved scores in animacy, likeability and helpfulness.

[1] https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/

More recent examples of researchers using the Godspeed questionnaire are Rifinski et al. [REF+20] and Terzioglu et al. [TMS20]. Even though they investigated different topics they both used a robot which was very mechanical looking, in the case of Terzioglu et al. [TMS20] it was a mechanical robot arm (Universal Robots UR5 cobot), while in the case of Rifinski et al. [REF+20] it was a non-humanoid, 2 degree-of-freedom robot called Kip. Terzioglu et al. [TMS20] tried to make the robot arm more human-like by using animation techniques and giving it clothing items and a breathing motion. They found that especially the breathing improved the perceived anthropomorphism of the robot. Rifinski et al. [REF+20] used their robot in a conversation between two humans as a third party participant. The robot showed responsive behaviour either towards the speaking or towards the listening person, or towards neither of them. The researchers found that the robot was perceived as more intelligent when being responsive to the speaking person.

Andriella et al. [ASF+20] used the robot NAO in combination with the Godspeed questionnaire to study how a robot's personality, expressed by verbal and non-verbal cues, influenced a human's perception of the robot. They had the robot help the participants with a memory game, the robot being introverted in one scenario and extroverted in another. They found that when the robot was behaving in an extroverted way, the extroversion and agreeableness score increased while the number of mistakes decreased. This shows how the Godspeed questionnaire can be used effectively to compare different types of non-verbal cues, in this case cues that demonstrate intro- or extroversion.

2.1.3 Studies Involving Videos of Robots

Woods et al. [WWKD06] found that video studies are a viable option when an interaction is limited. In addition, they state that timing and synchronisation of movement are crucial for an immersive experience. The fact that there is still a difference between watching a robot interact with a human in a video and interacting with the robot oneself remains. However, one advantage of video studies is that it is much easier to provide a consistent experience and therefore consistent data.

Walters et al. [WSD+08] used videos of robots as well as a live study to investigate how differences in the appearance of a robot can affect how the robot is perceived. They used two different study methods because they wanted to show that video studies were a relevant tool for the evaluation of robot perception. They argue that video studies also had some advantages over live studies because it allowed for a consistent experience in different scenarios. The Peoplebot was used with three different heads, arms and voices, to create three distinct appearances. They found that the majority of participants preferred the human-like appearance of the robot and that changing the appearance changed the participant's perception of the robot's personality. They also found that in a scenario with little HRI, the video study proved a valid method compared to the live study. Zecca et al. [ZME+09] used videos of the robot KOBIAN displaying different emotions for their study. The participants had to chose which emotion the robot in the video was displaying and the researchers managed to achieve a high recognition rating. Takayama et al. [TDJ11] used videos of a virtual robot when studying the effect of using

animation principles vs. not using them on readability and perception of a robot. They found some support for their hypothesis that the animation principle of *forethought* would improve readability and perception of the robot.

Li et al. [LJN15] used the NAO robot for two studies in which the participants were shown videos of the robot talking to a human. In the first study, they focused on one party, either the robot or the human, displaying authority over the other. In the second study, they made one party mirror the other party's body movements. This way they were able to display either the human or the robot as dominant or submissive and they found that the human party was trusted the same in either position. The robot was trusted less when it displayed dominance and even less when the human displayed submissive behaviour. This shows how videos can be used effectively to study a human's perception of a conversation between a robot and a human.

2.2 Non-Verbal Cues

When trying to understand non-verbal cues, it is useful to first take a look at non-verbal communication. Non-verbal communication can be explained as the sending and receiving of non-verbal cues, which can be spontaneous or deliberate [HHM19]. These non-verbal cues include, but are not limited to, the sender's gender, age and dress, but also their facial expression, posture and movements of body parts [HHM19; BBE+20]. This work will focus on the movements of different body parts, including gestures in the form of arm and hand movements. Non-verbal cues in humans have been studied for a long time: in 1872 Darwin [Dar15] already documented different facial expressions of humans and animals. But more recently, researchers have tried to adapt this knowledge for their studies on non-verbal cues for robots.

2.2.1 Non-Verbal Cues for Robots

Non-verbal cues for robots include posture, proximity and touch as well as movement or position of any body part including facial expressions and gaze [BBE+20]. All of the current state-of-the-art robots are limited in at least one of these areas, while most are limited in many of them. Some examples include the already mentioned humanoid robots Wakamaru, NAO, Pepper and Asimo, which do not have the ability to display facial expressions. While Asimo has the ability to move individual fingers, NAO and Pepper can only make a motion with all of their fingers at once, such as opening and closing their hands, and Wakamaru cannot move its fingers at all, the only hand movement possible is a twist of the wrist. These are a few examples of how some of the expressions of non-verbal cues are limited for some robots.

> "Nonverbal cues are such an important aspect of human communication that being unable to produce and decipher them appropriately makes interaction quite challenging." [BBE+20, page 81]

Bartneck et al. [BBE+20] explain the importance of non-verbal cues for communication. While it is not necessary that all non-verbal cues for robots are exact replications of non-verbal cues used by humans, creating cues that are familiar to humans can have many benefits. For instance, one benefit of the recreation of human behaviour is that it can lead to an increase in the persuasiveness of the robot [CCM12] or the willingness and enjoyment of humans talking to the robot [FIPC15]. Another benefit is a possible improvement in the effectiveness of teamwork [TMS20].

Two non-verbal cues which may be less obvious but nevertheless important are mirroring and backchanneling. Mirroring is when one interaction partner mirrors the non-verbal behaviour of the other partner. Backchanneling is when one of the listening interaction partners indicates to the speaking partner that they are listening or understand what the speaker is saying. This is often done non-verbally with nodding or gestures or verbally with utterances such as “hmh” or “okay” [BBH20].

2.2.2 Design and Evaluation of Non-Verbal Cues for HRI

A number of studies [CCM12; LIIH13; XBHN14; CDFV18] have looked into non-verbal cues for robots. Each of them focuses on different cues or aspects and effects that these cues can have. Some studies combined non-verbal and vocal cues. One example is a study by Chidambaram et al. [CCM12] in which they used the Wakamaru robot to study how non-verbal and vocal cues influence the persuasiveness of the robot, robot intelligence and the satisfaction humans feel when interacting with the robot. The authors used a desert survival task, where participants had to rank items they would take on a desert island to survive. The robot tried to influence the participants by using the non-verbal cues proximity, gaze and gestures as well as vocal cues. They utilised 4 different scenarios: One where the robot was using no cues, one where it was using just non-verbal cues, one where it was using just vocal cues, and one where it was using both. When comparing vocal and non-verbal cues in isolation, they found that non-verbal cues, in contrast to vocal cues, did improve compliance and were also more effective in the persuasion of participants. This shows that participants reacted differently to the robot when it was using non-verbal cues such as gaze and gestures. Chidambaram et al. [CCM12] looked at how these cues affected participants’ behaviour.

Liu et al. [LIIH13] have done research on creating head nodding, tilting and gazing for human-robot speech interaction. To create these motions for their robots, they used human head nodding and tilting motions from a database, which they had created for a previous study [ILIH10]. They then investigated how these head motions were perceived when used by the robots Geminoid F and Robovie R2. They specifically looked at the perceived naturalness of the robot’s movement and found that the naturalness rating was the highest when all of their created motions were employed by the robots. This research is especially interesting, as their work was the only one found to describe the process of creating non-verbal cues in great enough detail to be replicated.

Xu et al. [XBHN14] used the Nao robot to investigate whether a change in a robot’s

mood could be recognised by the participants, whether the mood could be contagious and whether it affected the participants' performance. They did this by creating different movements for the robot, depending on its mood. These changes influenced vertical and horizontal head movement, finger rigidness, palm direction and movement speed. Their studies contained an imitation game that was played by participants with the robot. The robot moved its arms and the participants tried to imitate the movement. The movements of the robot were slightly altered depending on the mood it was displaying. They found that participants were able to distinguish between a robot's positive and negative mood. The mood also had an effect on the participants' mood and their performance.

Non-verbal and verbal cues can also be used as indicators for different types of personalities. A study about this was done by Craenen et al. [CDFV18]. They used the Big-Five personality trait model [SG96] to determine the participants' personalities. During their study, the participants observed different gestures with different stimuli performed by the robot Pepper and had to score them using the Godspeed questionnaire. Their results show that 15 out of their 30 participants displayed the *similarity-attraction effect*, which describes the tendency to rate the robot displaying a similar personality to ones own more favourably. Meanwhile, 9 out of 30 participants displayed the *complementary-attraction effect*, which means that they rated the robot conveying a different personality to their own more favourably.

Bliek et al. [BBH20] used the Pepper robot to find out if robots could trigger human backchanneling by using backchannel-inviting cues like gestures and pauses. They found that not only did these cues increase the human's backchanneling but also that human backchanneling is different when talking to a robot versus another human. They used stories told by the robot to create a setting where the robot was talking and using backchannel-inviting cues and the participant was listening. This shows how storytelling can be used as a tool to study human perception or subconscious reaction to a robot's non-verbal cues used during talking.

Bergmann et al. [BKE10] did not use a robot but instead the virtual human Max for their research on how gesturing behaviour influences user perception. They compared different methods of gesture generation and their baseline condition consisted of random gestures. Their evaluation included the virtual agent verbally describing something while using gestures that were different for each condition. The different conditions consisted of two different decision networks and three control variables: No gestures, random gestures and a combination of the two networks. They found that while both of their ways of creating gestures received high ratings in all of their categories, the condition where both of them were combined received lesser score. The baseline condition, or control variable, of random gestures was given low ratings in every category and even rated lowest out of all the conditions in comprehension, vividness, likeability and competence.

Babel et al. [BKM+21] used the robot NAO to study the influence of gaze on people's perception of the robot. In their setup, the robot used either a fixed or random gaze during a conversation with the participant that was lead by a dialog script. They found

that the fixed gaze, when used during the conversation, lead to an increase in perceived anthropomorphism and acceptance, compared to the random gaze.

2.3 Animation Techniques and Exaggeration

In 1995, Frank Thomas and Ollie Johnston, two animators at Disney, had their book *The Illusion of Life: Disney Animation* published. It included twelve principles of animation with which cartoon animators could make their animations appear more life-like. These twelve principles are: *squash and stretch*, *anticipation staging*, *straight ahead action and pose to pose*, *follow through and overlapping action*, *slow in and slow out*, *arc*, *secondary action*, *timing*, *exaggeration*, *solid drawing* and *appeal* [TJ95].

Although the animation principles were first invented to help pen-and-paper animators, they started to be used in other fields, like 3D animation and computer graphics [LR87]. After their successful implementation in these other domains, they started to be used by some researchers to create movements and behaviours for robots [STH18].

Research [STH18] has shown that among the most commonly used animation techniques when working with robots are *secondary action* and *straight ahead action and pose to pose*. Secondary action means that in addition to the main action the character is carrying out, like opening a door, a secondary action is added, like the wiping away of sweat, to indicate the character being nervous to open the door [TJ95]. *Straight ahead action and pose to pose* are two different ways of creating a movement. In *straight ahead action*, the movement is created from beginning to end. In *pose to pose*, the main poses of the movement are created and afterwards the poses in between are filled in.

A principle of animation which has been studied less frequently by scientists is the principle of *exaggeration*. A definition for the principle of *exaggeration* has been given by Lasseter and Rafael [LR87]:

> "The meaning of exaggeration is, in general, obvious. However, the principle of exaggeration in animation does not mean arbitrarily distorting shapes or objects or making an action more violent or unrealistic. The animator must go to the heart of anything or any idea and develop its essence, understanding the reason for it, so that the audience will also understand it. If a character is sad, make him sadder; if he is bright, make him shine; worried, make him fret; wild, make him frantic." [LR87, page 41]

As Lasseter and Rafael [LR87] mentioned, the meaning of *exaggeration* seems obvious at first, but they then go on to describe how it is supposed to be used in animation. Exaggerating something can be an intuitive or even artistic process [RP12] in which movements are made bigger or more extreme but it can also be a process of precise calculations [PLC15] or even algorithms [GT12]. For example, when Gielniak and Thomaz [GT12] used *exaggeration*, they created an algorithm to exaggerate motion for robots,

which manipulated the rotation of the robot's joints according to calculations about the given movement.

2.3.1 Animation Techniques in HRI

The use of animation techniques when creating non-verbal cues for robots has been explored by a number of researchers. The work of Takayama et al. [TDJ11] has already been mentioned in this section in the context of studies using videos of robots. In this subsection, the use of animation principles in their research to improve the readability of robots' non-verbal behaviour is highlighted. Takayama et al. [TDJ11] tested whether certain animation principles would increase the readability as well as how they affected a participant's perception of the robot. One of their readability testing methods was a keyword matching system, which tried to find specific key words in the participants' descriptions of the action of the robot. Their results were mixed but their keyword matching results varied more between tasks than between the different animation techniques. This might indicate a flaw with the keyword matching readability test system rather than the implementation of the animation principles. The work of Terzioglu et al. [TMS20] has also been mentioned before. They used the animation techniques of *secondary action*, *arc* and *appeal* and found that a breathing motion increased the life-likeness and that gazing behaviour can improve most interaction outcomes.

In their systematic literature review of animation techniques being used in HRI, Schulz et al. [STH18] found that the principle of *exaggeration* was rarely used. Schulz et al. [STH18] looked at the twelve principles of animation and all the studies they could find that used any of them in the context of HRI. They analysed 27 papers, two of which used the principle of *exaggeration* [GT12; PLC15].

In total, only three studies could be found which used the design principle of *exaggeration* in the context of robots [GT12; PLC15; RP12]. As mentioned, Schulz et al. [STH18] found two studies which used the design principle of *exaggeration* in the context of non-verbal cues for robots. The first study found by Schulz et al. [STH18] was by Gielniak and Thomaz [GT12], who carried out an experiment using the SIMON robot and created non-verbal as well as exaggerated cues for it. They found that *exaggeration* was not only recognised by participants but also increased the engagement and perceived entertainment value. The second study was by Park et al. [PLC15], who focused on facial expressions using the Kismet robot, which they had created themselves. They had the robot perform facial expressions for different emotions, which they then exaggerated. They did not compare the exaggerated and non-exaggerated version but rather just used *exaggeration* as a step in the generation of facial expressions.

In addition to the papers identified by Schulz et al. [STH18], two other studies could be found that look at the principle of *exaggeration*. Firstly, Ribeiro and Paiva [RP12] used the EMYS robot and concentrated on facial expressions. They used multiple principles of animation including *exaggeration* to increase readability of different emotions and evaluated their results using an online survey and videos of the robot. In their survey,

they used three different intensities of their facial expressions and found that all of them had high recognition ratings. They did not, however, statistically analyse the difference between the three types of intensities, therefore no conclusion can be made from the data on whether the use of animation principles impacted readability. Secondly, Atkinson et al. [ADGY04], who looked at the recognition of emotional postures that were performed by humans, and exaggerated versions of them. The last study did not include robots but rather videos of humans. However, they did also include a condition where only the joints of the humans were highlighted by light strips and the rest of the video was dark. In this condition, the focus is more on the structure of the movement than on the human as such. As has already been mentioned in Section 1.1, they found that *exaggeration* improved the readability of emotional postures for all of their five emotions except sadness. However, these results were only true for when they used the light strips. When they used the full image of the human, the results remained true for only two instead of four of the five emotions.

CHAPTER 3

Methodology

After having discussed the related work, we will now focus on the methodology employed for the creation of non-verbal cues for the robot Pepper. First, general information about Pepper will be provided and in a second step, the selection of cues chosen for this paper (nodding, head shaking, leaning and gestures accompanying speech) will be presented in detail. This is followed by a detailed look at the implementation of these cues, the program that was used and the process of creating the individual cues. The different cues were validated and evaluated using two different online surveys and four semi-structured interviews. Next, the setup of the online surveys is described. As the surveys include videos of Pepper, these videos needed to be created first. Therefore, Section 3.4 starts with a description of tools and processes used for the creation of the videos. This is followed by an explanation of how the server was set up. Next, the consent forms and general questions, which were used in both surveys, are discussed. The first online survey, which was part of the validation of the cues and therefore called validation survey, is described next. This is followed by information about the second online survey, which was part of the evaluation of the cues and therefore called evaluation survey. Section 3.5 provides information about the interviews which were done as part of the evaluation. For an outline of the entire process see Figure 3.1.

Figure 3.1: The process for the creation and evaluation of non-verbal cues for the robot Pepper.

3.1 The Pepper Robot

The Pepper robot, which can be seen in Figure 3.2, is a socially interactive humanoid robot developed by SoftBank Robotics [Robf]. It is 1.2 m tall and has 17 joints which enable it to display a large range of movements for body language. The robot also has the ability to talk through loudspeakers located on both sides of its head but is unable to change its facial expression or move its mouth while speaking. Although the robot is humanoid, it was designed to not resemble humans too closely in appearance as to avoid an effect which can create a feeling of unease in humans when virtual characters or robots look very similar to but not exactly like actual humans [PG18].

Figure 3.2: A photo of Pepper [Robe].

3.2 Selection of Non-Verbal Cues

The creation of non-verbal cues started with the selection of the non-verbal cues that would be designed in the study. For this, studies which have used non-verbal cues for robots were examined and the different types of cues were laid out. The cues which were selected to be investigated in this study are: nodding, head shaking, leaning and gestures accompanying speech. Next, the methods of creating the nodding, head shaking and leaning needed to be selected. The method of creating gestures accompanying speech is different to the one used for nodding, head shaking and leaning and will be explained in Section 3.3.4. The creators of the Pepper robot, SoftbankRobotics [Robf], provide the program Choregraphe for the creation of movements and behaviours for their robots, including a library of pre-implemented movements for Pepper. The program and how it was used as part of this project will be explained in more detail in Section 3.3. As there are already pre-implemented movements, including non-verbal cues, in Choregraphe, some of them were selected to be used for this work. This had the advantage that these movements would be accessible to anyone with access to a Pepper robot, which makes them easily reproducible for other researchers. Additionally, from-scratch versions of the non-verbal cues were created in order to compare their readability with the Choregraphe's cues. These from-scratch cues were created by us, with inspiration from human motion

and different papers, which will be credited when describing the creation of the respective cues. As many of the studies that we looked at did not include enough information to facilitate replication. The nodding was created after the cue from Liu et al. [LIIH13] and the head shake and leaning were created using inspiration from Mutlu [Mut11] and Rifinski et al. [REF+20]. This resulted in 2 x 2 (Choregraphe / From Scratch x Non-Exaggerated / Exaggerated) versions for each of the three cues (nodding, head shaking and leaning). An overview of this can be found in Table 3.1.

Cue	Choregraphe	Liu et al. [LIIH13]	From Scratch
Nod	X	X	
Nod Exaggerated	X	X	
Head Shake	X		X
Head Shake Exaggerated	X		X
Lean	X		X
Lean Exaggerated			X X

Table 3.1: The 12 non-verbal cues that were created (each represented by an X). The table shows the type of non-verbal cue on the left and its origin on the top. The colour of the X indicates whether the cue was created after or taken from Choregraphe (blue) or created after or taken from literature (orange).

Each X in the table represents one of these 12 cues, the blue ones represent the cues taken from Choregraphe's library and the orange ones represent the ones created by us. The two Xs in the second column stand for the nod, which was created after the model of Liu et al. [LIIH13]. The blue X in the third column is there because there was no exaggerated lean in Choregraphe's library of pre-implemented non-verbal cues. Therefore, this cue had to be created by exaggerating the non-exaggerated lean from Choregraphe. So even though both exaggerated leans were created from scratch, they are different because they are each an exaggeration of a different lean, one from Choregraphe (the blue X) and one created from scratch (the orange X).

3.3 Implementation

In this section, the implementation of the non-verbal cues for the robot Pepper is explained. It also features a description of how the four types of non-verbal cues, nodding, head shaking, leaning and gestures, were created and exaggerated. As mentioned previously, the cues that were implemented for this thesis were created using the program Choregraphe. The program is based on a visual programming language which uses boxes (that contain code modules) which are connected with lines to determine the process order. These boxes can be accessed through the box library, which contains all of Choregraphe's pre-implemented modules (or boxes). The feature which was used most was Choregraphe's timeline, which is one of the modules that can visualise different movements of the robot's

limbs with Bezier curves. Bezier curves can be used to visualise the movement of the limbs and are drawn in a grid (Cartesian coordinate system) of angle (in degrees) and time (in Choregraphe's own time unit). These time units do not match any standardised units but were invented by Softbank Robotics. To find out how the Choregraphe time units translate to standard seconds, a short Python program was written within Choregraphe to measure the time which elapsed during a set amount of Choregraphe time units.

As mentioned before, 12 different non-verbal cues were created, including four nods, four head shakes and four leans. For each cue, four different versions were created: a non-exaggerated and exaggerated version created in Choregraphe and a non-exaggerated and exaggerated version created from scratch (see Table 3.1). For the creation of the nodding and head shaking cues, Pepper's joints, which can be seen in Figure 3.3, were used. The hip and knee joints were used for the leaning motions and the joints in the arms, including the shoulders and hands, were used for the gestures during speech. A more in depth look into the different joints will be provided in the respective sections: the head joints will be discussed in Sections 3.3.1 and 3.3.2 and the hip and knee joints in Section 3.3.3.

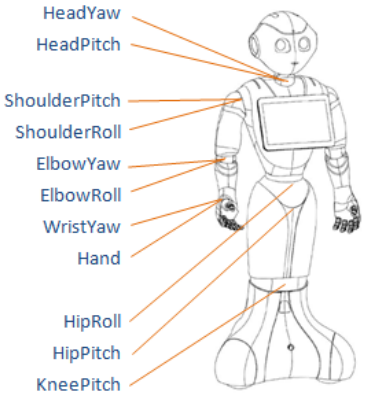


Figure 3.3: The joints of the robot Pepper [PG18].

3.3.1 Nodding

To create the nodding movement, the robot's *HeadPitch* was changed: A scheme of the movement capabilities of the motors in the robot's head can be found in Figure 3.4. The figure shows how Pepper's head can tilt forward 36.5 degrees and backwards -40.5 degrees.

Choregraphe already has some pre-implemented animations for Pepper in the box library, some of which include a nodding motion. They were extracted from the animation, which is called *OfferBothHands_ HeadNod_LeanLeft_01*, while the other parts of the animation were not used, as to not distract from the nodding motion. The Bezier curve for the Choregraphe nod can be seen on the left in Figure 3.5. The head first moves back, to almost -15 degrees, and then forward to less than 5 degrees. Afterwards it moves back

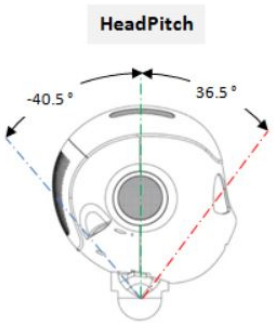


Figure 3.4: Head joint for the HeadPitch of the robot Pepper [Robd].

to less than -10 degrees. This sequence creates a nodding movement, over the course of 40 of Choregraphe's time units.

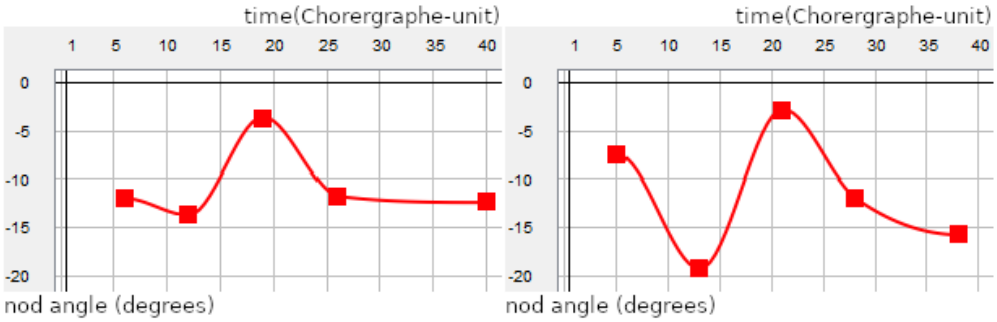


Figure 3.5: The Bezier curves for the head nods from Choregraphe's library. Left: Choregraphe's nod. Right: Choregraphe's exaggerated nod. Both show the change of the *HeadPitch* (red) in degrees, over time.

For the exaggerated nod, another predefined animation from Choregraphe was used, the *StrongNodArmsUpAndDownLeaningBack_01*. All movements of other body parts were removed to create the exaggerated head nod. This resulted in the red Bezier curve which can be seen on the right in Figure 3.5. The exaggerated head nod from Choregraphe is not an exact copy of the non-exaggerated version, but rather a slightly different curve. This is because Choregraphe's nod and Choregraphe's exaggerated nod were used and they are not exactly the same. The decision was made to use them regardless, as they were both pre-implemented by Choregraphe and therefore offer easy reproducibility and access by other researchers.

For the from-scratch nod, Liu et al.'s [LIIH13] nod was adapted for Pepper. In their study, they used the robots Geminoid F and Robovie R2 and created the nod for them by analysing human head nods. They created a representative motion shape for the nod which can be seen on the left in Figure 3.6.

For the exaggeration of our nod that was inspired by the model of Liu et al. [LIIH13],

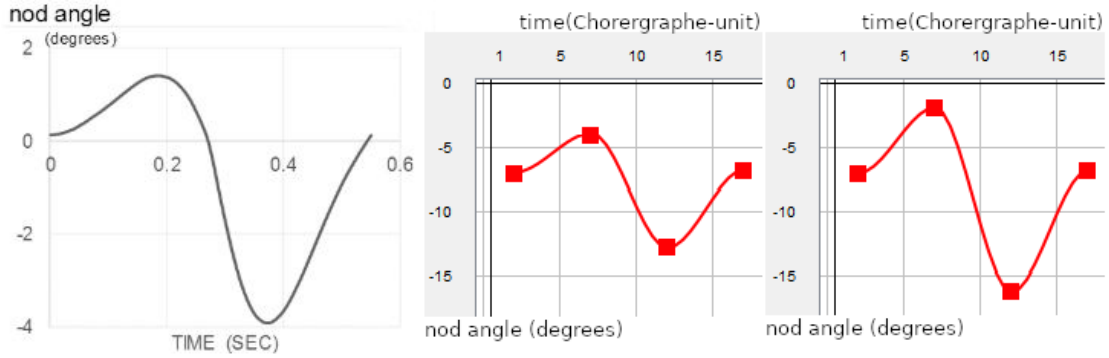


Figure 3.6: The Bezier curves for the head nods created by and after Liu et al. [LIIH13]. Left: nod by Liu et al. [LIIH13] Middle: from-scratch nod created for this thesis, based on Liu et al. [LIIH13] Right: exaggerated from-scratch nod created for this thesis, based on Liu et al. [LIIH13]. All show the change of the *HeadPitch* (grey / red) in degrees, over time.

a similar approach to Ribeiro and Paiva [RP12] was used. For their study, important features of the movement were exaggerated based on the movements of puppets, since they are closer to robots in terms of movement. Atkinson et al. [ADGY04] used actors who were told to exaggerate their movements. Both of these approaches served as inspiration for the exaggeration we used in our work. To create the exaggerated movement, the two peaks of the original curve, which can be seen in the middle of Figure 3.6, were both made more extreme. The results can be seen in the same figure on the right.

3.3.2 Head Shaking

In contrast to the nod, the head shake uses the *HeadYaw*, which has a maximum of 119 degrees to either side and can be seen in Figure 3.7. The head shake from Choregraphe was taken from the predefined animation *RightArmUpAndDownWithBump _HeadShake_01* and its Bezier curves can be seen on the left in Figure 3.8. This figure also shows that Choregraphe's head shake, unlike the nod, repeats multiple times. As with the other cues taken from Choregraphe, the movement of the other body parts was removed.

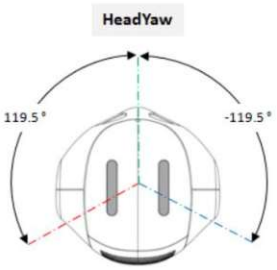


Figure 3.7: The head joint for the HeadYaw of the robot Pepper [Robd].

Choregraphe's exaggerated head shake was taken from the predefined animation *WideHorizontalRightArm _LeanLeft_HeadShake_01*. It can be seen on the right in Figure 3.8. This is, as with the exaggerated nod from Choregraphe, not an exact reproduction of the non-exaggerated version with an added exaggeration, but a different movement that represents an exaggerated head shake. For this cue, as opposed to the nod, not all the other movements were removed because there was a change in *HeadPitch* as well as *HeadYaw* that was considered crucial for the expression of the head shake. This change in *HeadPitch* was not present in Choregraphe's non-exaggerated head shake.

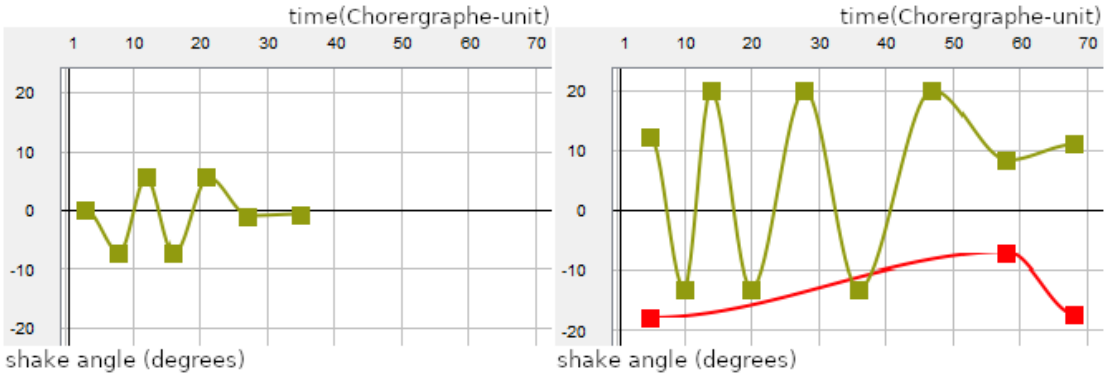


Figure 3.8: The Bezier curves for the head shakes from Choregraphe's library. Left: Choregraphe's head shake. Right: Choregraphe's exaggerated head shake. Both show the change of the *HeadYaw* (green) in degrees, over time. The right image also shows the change of the *HeadPitch* (red).

Since no publications were found in which a head shake for a robot was described in enough detail to be used in this study, the head shake was created from scratch. It started with a simple left to right movement of the head, similar to the head shake from Choregraphe (see Figure 3.8) Fujie et al. [FEN+04] created their head shake by repeating such a movement. Thus, for the final head shake, the movement was repeated to create a left-right-left-right movement of the head. The final Bezier curves of the head shake can be seen in Figure 3.9.

For the creation of the exaggerated from-scratch head shake, the same methods that were employed for the other exaggerations were used. This time, the movement was again made bigger by adding to the extremes of the movement. The amount of degrees was doubled in both extremes. A depiction can be found on the right in Figure 3.9.

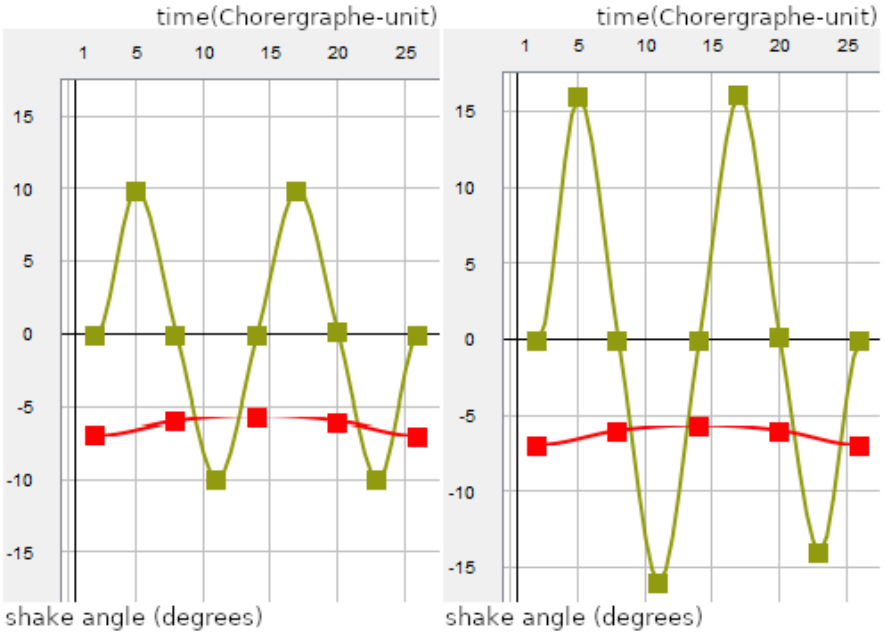


Figure 3.9: The Bezier curves for the head shakes created from scratch. Left: from-scratch head shake. Right: from-scratch exaggerated head shake. Both show the change of the *HeadPitch* (red) and the *HeadYaw* (green) in degrees, over time.

3.3.3 Leaning

Pepper's ability to move its hip and knee can be observed in Figure 3.10. To create the lean, the *HipPitch*, the *KneePitch*, and the *HeadPitch* were used. The HeadPitch is included in order for the robot to be able to keep eye contact with the person in front of them. The head needs to move up during the forward leaning motion, in order for the robot to be able to look forward. If only the hip and knee moved, the robot would look downwards at the end of the movement.

The *HipPitch* and *KneePitch*, in contrast to other motors, are closely connected to each other, meaning when one is moved the other automatically moves as well. This is important to ensure that the robot does not fall over or jeopardise its balance in any way. Therefore, the positions of the knee and hip are usually somewhat different, even if the same values are selected, because they automatically get adjusted. The amount of adjustment can vary a bit. This must be kept in mind when designing the leaning motion.

The head's and legs' (which include the hip and the knee) movement from the predefined *GoToStance_Exclamation _LeanBack* animation from Choregraphe was used for the lean. The curves can be seen on the left in Figure 3.11. The lean is a lean towards the front. The fact that the name in Choregraphe includes the words "LeanBack" is most likely because the hip has to lean back for the body to lean forward. This results in the

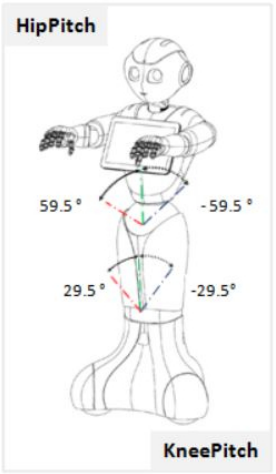


Figure 3.10: Hip and knee joints of the robot Pepper [Robd].

lean forward being called “LeanBack” and vice versa.

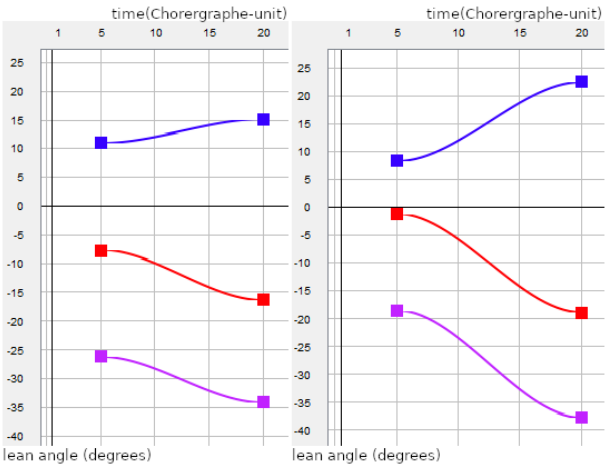


Figure 3.11: The Bezier curves for the lean from Choregraphe’s library. Left: Choregraphe’s lean. Right: Choregraphe’s exaggerated lean. Both show the change of the *KneePitch* (blue), the *HeadPitch* (red) and the *HipRoll* (purple) in degrees, over time.

There is no exaggerated lean in any of the default animations from Choregraphe. To be able to use an exaggerated lean to compare to the from-scratch lean, the lean from Choregraphe was exaggerated the way the from-scratch motions were exaggerated. The resulting curves can be seen on the right in Figure 3.11. The curves were made more extreme by adding or subtracting around 5 degrees to each angle. This, like in the lean from Choregraphe, included *HipPitch*, *KneePitch* and *HeadPitch*.

The lean was created by lowering the *HipPitch*, which results in the robot leaning forward. The Bezier curves for the head and hip movement of the created lean can be found on the left in Figure 3.12. But it is important to note that Pepper automatically adjusts the angle of the knee whenever the hip pitch is changed so that Pepper will not fall over when it leans forward or backwards. This means that even though the Bezier curves of the from-scratch lean only show a change in the *HipPitch*, automatic changes are made to the *KneePitch*.

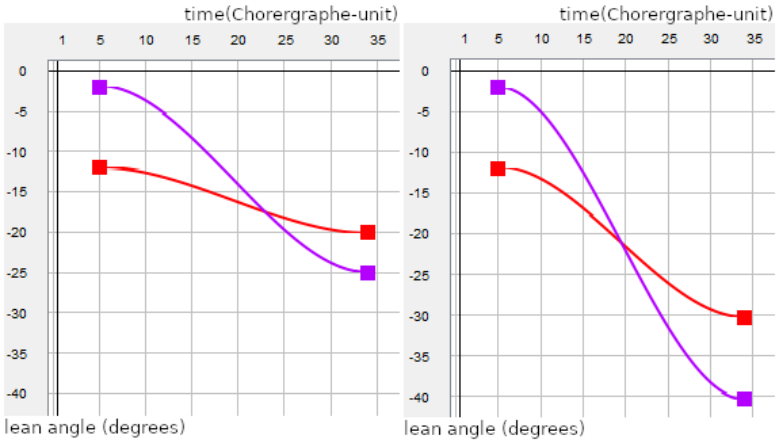


Figure 3.12: The Bezier curves for the lean created from scratch. Left: from-scratch lean. Right: from-scratch exaggerated lean. Both show the change of the *HeadPitch* (red) and the *HipRoll* (purple) in degrees, over time.

The exaggeration of the created lean was done the same way the other exaggerations were done. The curves were pulled into the minus direction, leading to an increased forward movement of the upper body and an increased upward movement of the head. The resulting curves can be seen on the right in Figure 3.12.

3.3.4 Gestures to Accompany Speech

Additionally to the non-verbal cues nodding, head shaking and leaning, gestures to accompany speech were created. Choregraphe has multiple options when it comes to Pepper using gestures while talking. Choregraphe's box library contains the boxes *Animated Say* and *Animated Say Text* which both include the option for a *speaking movement mode*. There are three different modes: Disabled, random and contextual. While Pepper will not be doing any gestures when the disabled mode is selected, when the random mode it selected it will perform random gestures with its speech. These random gestures are based on the box library's pre-implemented animations. If the contextual mode is selected, Pepper performs gestures with context of the word it is saying.

In order to create contextual gestures for Pepper, the following pre-implemented animations from Choregraphe's box library were used: *BodyTalk 1, 2, 3, 4, 11, 13*, *Listening 4*,

Sad 2 and *Sorry 1*. The contextual gestures could theoretically be used with code such as:

```
^start(BodyTalk_1) A Dog is walking home ^start(Listening_4)
```

Here, the codes *start(BodyTalk_1)* and *start(Listening_4)* refer to the start of the animation *BodyTalk_1 / Listening_4* and the sentence *A Dog is walking home* describes the words the robots is saying while displaying the animation. The rest of the text is said by the robot and can be found in Appendix A. However, this code did not work in our tests, which could be in part because the Linux version of Choregraphe was used. The problem was that Pepper was not using the contextual non-verbal cues, even though the mode had been switched to the contextual. Therefore, the author of this thesis had to create a custom sequence of *Animated Say* and *Text* boxes. For the random gestures, the *Animated Say Text* box on the random setting worked and Pepper produced random gestures to its speech. The next step was the exaggeration of the contextual gestures. This was done by taking each of the pre-implemented animations from Choregraphe's box library that were used for the contextual speech and exaggerating them one by one. The methods for exaggeration were the same as the ones used for the exaggeration of the other cues, nodding, head shaking and leaning. An example of a gesture and its exaggerated version can be found in Figure 3.13.

Additionally, it should be mentioned that the *Animated Say* and *Animated Say Text* boxes, in any mode, could not be tested on the virtual robot, which can be used in Choregraphe if the physical robot is not available, to emulate the robot's movements and behaviours. Therefore all the testing and implementing of these features needed to be done with the physical robot.

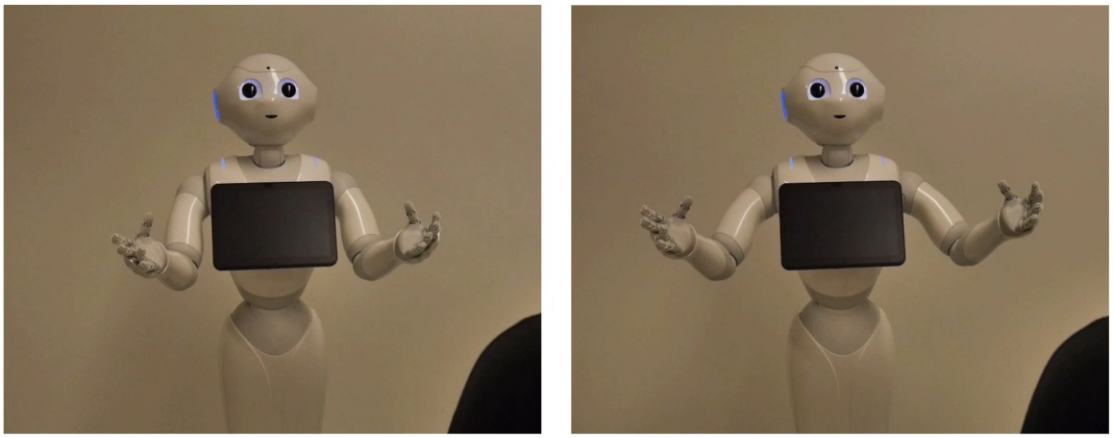


Figure 3.13: Pepper using a gesture while talking. Left: non-exaggerated gesture. Right: exaggerated gesture.

3.4 Online Surveys

Two online surveys were carried out over the course of this work. Both of these surveys used videos and were set up using LimeSurvey [Gmb]. This section starts by explaining how the videos for the surveys were created and how the servers were set up. This is followed by a description of the consent form and general questions which were used in both surveys. Next, the validation survey and evaluation survey are explained in detail, including information about the participants, videos and questions of the respective survey.

3.4.1 Creating the Videos

A Panasonic DMC-GX7 camera was used to record video and audio for the videos used in the surveys. Additionally, audio was recorded with a Motorola Moto G6 Phone. The additional audio recording was used in the final videos, the audio recording from the camera was only used to synchronise the additionally recorded audio with the video. The audio was recorded separately because it was possible to put the microphone very close to the robot and capture the sound more clearly than with the microphone from the video camera, which was further away. Before the audio was used in the videos, the background noise was reduced in the program Audacity [Aud], using the built-in noise reduction effect. Afterwards the video and audio tracks were synchronised and the videos were cut to the right length in the program Hitfilm Express [FXh]. To support as many browsers as possible, the videos were provided in different formats, for the transformation of the original mp4 format, the program VLC player [LAN] was used. The videos were transformed into the format ogg and webM because these formats are supported by the most commonly used browsers. The videos were then embedded into the survey using html5. An example frame from one of the videos can be found in Figure 3.14.

Since the videos for the evaluation online survey included speech, they needed subtitles: To create those, the program Handbreak [Tea] was used. In order to ensure that all the videos where the person was speaking to the robot sounded the same, the same voice audio file was used for every video. This was then combined with the background noise audio, in order to still have the original sounds of the motors in each video.

3.4.2 Server Setup, LimeSurvey and Data Proctection

The surveys were created using LimeSurvey, an online survey tool which can be set up on a private server. This is relevant, because this way the access to the data can be controlled. LimeSurvey was installed on a personal server using the FreeBSD ports [Fre], which is a repository of binaries for servers running on the FreeBSD operating system. LimeSurvey is an open-source software which ensures that no secret back doors, which would enable third party to access the data, are included. Since it is installed on a personal server, it ensures that the data is only accessible for the researcher.

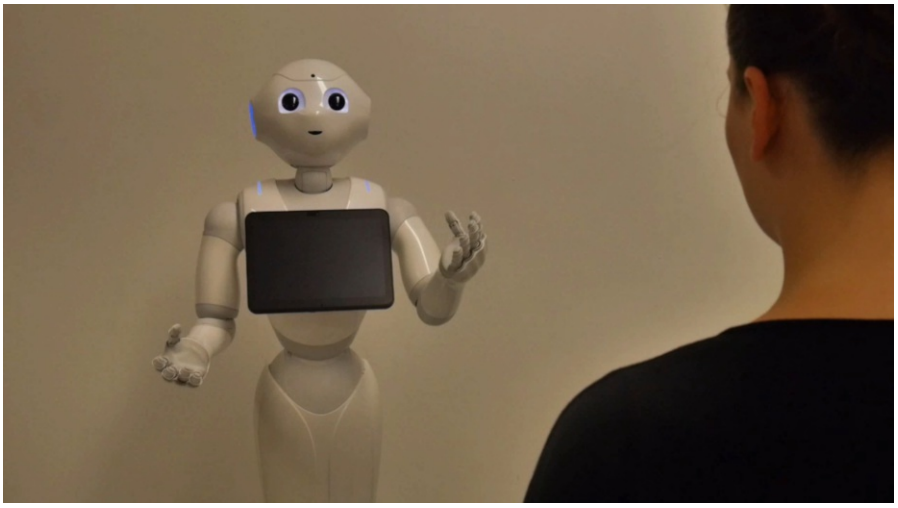


Figure 3.14: A frame from one of the videos created for the evaluation survey showing Pepper while it is talking to a person, using gestures.

For the videos for the evaluation survey, LimeSurvey was used again, but in contrast to the validation survey, the private server was unable to load the videos fast enough, as the videos for the evaluation survey were longer and therefore needed more resources to load without stalling. To be able to still ensure data protection, a server was set up at the Vienna University of Technology (TU Wien) and LimeSurvey was installed there.

When configuring both servers, some upload file sizes needed to be increased, since videos needed to be uploaded. The limit for the PHP variables *post_max_size*, *upload_max_filesize* and *upload_max_filesize*, as well as the *client_mac_body_size* was increased. Also the file extensions .mp4, .ogv and .webM needed to be included in the variables *allowedresourcesuploads* und *allowedfileuploads*.

To further ensure data protection, the survey only collected data anonymously. This is an option which can be selected within the LimeSurvey program to ensure that the participants cannot be linked to their data. It has the advantage that no one can identify participants because no cookies or identification tokens are used, and the IP addresses of the participants are not collected. Disadvantages are that it cannot be ensured that a participant did not participate in the survey more than once and that partially answered surveys cannot be saved and continued at a different time.

3.4.3 Consent Form and General Questions

Before the survey, the participants had to confirm that they were 18 years old or older, to ensure that they fit the target group, and that they had read, understood and agreed with the points on the consent form page. These points contained the consent and agreement to the voluntary participation, possible withdrawal at any point during the survey as

well as information on the anonymous nature of the collected data and on where and how the data would be used.

The online survey proceeded with general questions about the participant: The country they spent most of their lives in, their age, gender and the highest level of education as well as the field of education. Additional questions were, whether they were a student and rating their experience with robots on a scale from 1 to 5. The entire survey including the consent form and general question were available in German and English. The exact wording of the consent form and the general questions (in both languages) can be found in Appendix A.

3.4.4 Validation Survey

The validation survey consisted of 12 videos, which are each about 3 seconds long. The videos each show the robot Pepper performing one of the cues: nodding, head shaking or leaning, once non-exaggerated and once in exaggerated form and both, a version from Choregraphe and one created from scratch. The gestures to accompany speech were not part of the validation, because the validation survey was done to find out if the created cues nodding, head shaking and leaning were readable to humans. This was not necessary for the gestures to speech, as they did not have to rely on readability because the spoken words already provided context and the gestures were there to accompany the speech, not to be understood on their own.

Participants

In the validation survey 39 people participated, 16 (41.03%) of which were female and 23 (58.97%) were male. All of them chose to answer the German version of the survey. Most participants (94.87%) said to have spent most of their lives in Austria. Only two (5.13%) stated that they spent most of their lives in Italy and Germany respectively. Out of the 39 participants 29 (74.36%) stated that they were students, while one person did not disclose whether they were a student or not. The mean age was 27.36 years with a standard deviation (SD) of 6.71. The mean value of experience with robots was $\eta = 2.49$ with $\sigma = 0.91$.

Videos and Questions

The order in which the videos appeared was randomised. After each of the 12 videos, there were two questions, one being a multiple choice question which asked what the robot was tying to convey. The possible answer options, which appeared in random order, were *Agreement*, *Disagreement*, *Attention / Interest*, *Defensiveness / Disinterest* and *Other*. The second question was an open-ended question where the participant was asked to describe the robot's body movements. The exact wording of the questions and possible answer options (in both languages) can be found in appendix A. The results were analysed using a Qui-Square test, as well as descriptive statistics.

3.4.5 Evaluation Survey

The evaluation survey consisted of 5 videos which were each around 30 seconds long. To evaluate and compare exaggerated versus non-exaggerated movements, a story telling scenario was used. The full story can be found in the Appendix A. Three of the videos contained Pepper telling a human a story while using gestures and the two others contained a human telling Pepper a story and Pepper reacting with non-verbal cues. There where three different versions of the videos in which Pepper was talking: one where it used random gestures, one with contextual gestures and the last one with exaggerated versions of the contextual gestures (see Table 3.2). For the videos in which Pepper was listening to a story, there were only two versions: one where it used the non-verbal cues nodding, head shaking and leaning and one where it was using the exaggerated versions of these cues (see Table 3.2). The nod, head shake and lean that were used in the videos were chosen from the ones created for the validation survey. Initially, it was planned to chose the version (either Choregraphe or from-scratch) of the cue with the highest readability for the evaluation. But the data from the validation showed that there was no statistically significant difference in readability between the cues from Choregraphe and the ones created from scratch (for more detailed information see Chapter 4). Therefore, the versions of the cues had to be chosen by the author of this thesis. In order to keep one cue type from every type of creation (Choregraphe, Liu et al. [LIIH13] and from scratch), the nod chosen was the one created after the one from Liu et al. [LIIH13]. The chosen head shake was the one from Choregraphe and the chosen lean was the one created from scratch.

Video Content	Random	Contextual Non-Exaggerated	Contextual Exaggerated
Robot Talking	X	X	X
Robot Listening		X	X

Table 3.2: The allocation of cues in the 5 videos of the evaluation survey. Each X represents a video. The table shows the video's content, with the left representing the action of the video and the top showing how the action was performed.

Participants

56 people participated in the online evaluation survey, but two of the participants needed to be excluded due to not agreeing to the consent form. Out of the 54 valid responses three (5.56%) did the survey in English, and 51 (94.44%) used the German version. All but four (7.41%) people stated that the country they spent most of their lives in was Austria. The other four countries were the United States of America, Bulgaria, Bosnia and Herzegovina and Romania. The mean age was 31.93, with an SD of 12.26. Concerning the gender of the participants, 50% of the participants were male, 44.44% were female, two (3.7%) were non-binary and one (1.85%) did not want to disclose their

gender. As for education, 50% were students at the time of the survey and 68.51% had received higher education, such as a university degree, the fields were mostly computer science (37.84%), but also included humanities, economy, education and health care. When rating their experience with robots on a scale from 1 to 5, the mean of the answers was 2.8, with an SD of 0.92.

Videos and Questions

The order in which the videos appeared in the evaluation survey was randomised. After each video, the participants had to answer the Godspeed questionnaire by Bartneck et al. [BKCZ09]. As is recommended by Bartneck et al. [BKCZ09], the separation of the five categories was removed, and the order of the words was randomised. The exact words from the Godspeed questionnaire were used and for the German version of the survey the official German translation was used, which can be found on Bartneck et al.'s [BKCZ09] website[1]. The resulting data was analysed using a repeated measures analysis of variance (ANOVA) and a paired t-test.

3.5 Interviews

After the evaluation survey, semi-structured interviews were conducted as a part of the evaluation process. The interviews were done with four interviewees who previously participated in the survey, in order to gain more insight into the interviewees' experiences of watching the videos.

The four interviewees were chosen so as to include two female and two male interviewees. Each gender group was also divided into people from the field of computer science and people with a background in social sciences, who were less versed in technical sciences. All interviewees were students because 74.36% of participants in of the validation survey were students and the participants of the evaluation survey were expected to be similarly distributed. The goal was to pick interviewees which would represent the participants of the survey well, so as to get a good insight into the participants' point of view. This is also the reason why people between the ages of 26 and 27 were chosen for the interview, since the average age of participants in the validation survey was 27.

The interview guidelines (see Appendix A) were modelled after the Witzel's [Wit85] recommendations in order to learn more about views of the participants on the robot's non-verbal cues. Witzel [Wit85] suggests that the interviewer uses guidelines during the interview which contain ideas for questions about the each topic that will be covered in the interview. The open ended questions were created to answer the research questions, see Chapter 1. The goal was to let the participant recollect and explain their experience with as little interference from the researcher as possible.

The interviews were conducted via Zoom or in person in a quiet place with very little distractions. Before the interview, the participant was asked to read and sign a consent

[1] https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/

form (see Appendix A). During the interview, only the interviewer and interviewee were present and the interview was audio recorded so that it could be transcribed and analysed later.

The interview transcripts were pseudonymised with letters and numbers and analysed using the thematic analysis method by Braun and Clarke [BC12]. The method was used in order to find common themes among the interviewees. The goal was to find out more detailed information about the feelings and thought of the participants concerning the robot.

3.5.1 The Thematic Analysis

A thematic analysis as proposed by Braun and Clarke [BC12] is carried out in six phases, but beforehand the type of approach that will be used has to be selected. For this study, an inductive analysis approach was chosen, which means that the themes are not predetermined but rather emerge from and therefore are determined by the data. This approach was chosen because the intention behind this research is to follow where the data leads without too much initial interference, in order to find out what people's experiences are without presenting them with preconceived ideas. The approach is also semantic rather than latent, which means that instead of assumptions and interpretations of the data, the explicit content of the data is analysed.

Phase 1: Familiarisation

This phase is dedicated to the researcher's familiarisation with the data. This was achieved in this study by the researcher conducting the interviews, transcribing the data and reading the transcripts. This is an important step to identify any patterns that emerge and to be able to get an overview over the data.

Phase 2: Coding

In this step, the researcher goes through each transcript and codes sections of text labels, so called "codes". These sections of text are mostly phrases and parts of sentences, to which a code, which describes these phases, is assigned. This coding process is done for every interview transcript.

Phase 3: Generating Themes

In this phase, we use the codes from all the transcripts to identify groups and patterns within the data, which leads us to the themes which we are trying to create. Themes often include multiple codes and are therefore usually broader than codes. This is also the point where some codes will be discarded because they are not relevant enough.

Phase 4: Reviewing Themes and Phase 5: Defining and Naming Themes

During these steps, the themes are compared to the data set and it is determined whether they represent the data well. Then the themes can be split up or combined as well as renamed or even discarded.

Phase 6: Writing Down Results

The data has been collected and encoded, themes have been established and reviewed. Finally, the results can be written down and visualised. For this study, a sunburst diagram has been chosen as the visualisation technique. The colour indicates the theme and the size indicates the frequency of appearance of that theme or sub-theme. The results of the thematic analysis for this study can be found in Chapter 4.

CHAPTER 4

Results

In this chapter, the results obtained from the surveys and interviews along with the employed statistical methods are presented. Firstly, the results from the validation survey, which was concerned with design and readability of non-verbal cues, are laid out. Then the results from the evaluation survey, which looked at human perception of the robot, measured in anthropomorphism, animacy, likeability, perceived intelligence and perceived safety, are discussed and illustrated in tables and charts. The section about the evaluation survey results is separated into the results for the videos in which the robot is talking and the ones where the robot is listening. Lastly, quantitative data, collected through open-ended questions and semi-structured interviews, is analysed.

4.1 Design and Readability of Non-Verbal Cues

As discussed in Chapter 3, the three different types of non-verbal cues for Pepper, nod, lean and head shake (abbr. Shake), were created in three different ways: 1) Using cues from the Choregraphe library (abbr. Chore); 2) Using cues created after Liu et al. [LIIH13] (abbr. Liu); and 3) Cues created from scratch inspired by literature and human motion (abbr. FromScratch). For each of these cues a non-exaggerated and an exaggerated version (abbr. Ex) were created.

For the analysis of the data from the validation survey, the program SPSS Statistics [IBM] (SPSS) was used. SPSS is a statistics program which includes features for the calculation of different statistical analyses. To analyse the data from the validation survey, Pearson's chi-square test (chi-square test) was conducted.

Before the chi-square (χ^2) test can be explained, it is useful to first understand the data which we are trying to test. For this we will take a look at Table 4.1. In this table, the abbreviations for the different cues are listed on the left, and the five answer options (*Agree*, *Disagree*, *Interest*, *Disinterest* and *Other*) at the top. The numbers represent

the number of participants that chose the respective answer option for each of the cues. The *Recognised* column displays how many people recognised the body motion of the cue correctly (discerned from the free-text question). The *Answers* column displays the number of answers that were given (these differ because multiple answers could be selected).

The chi-square test is used for categorical data and binary data is categorical data that can only take one of exactly two values, for example 0 and 1. When taking a look at the results from the validation survey in Table 4.1 we can see that the data does not look binary at first glance. But if we consider that there was one answer option which can be considered correct for every cue, we can interpret our data as binary data, which can be seen as correct (1) and incorrect (0). For example, for the nod the answer *agree* will be considered correct (later we will also look at the answer option *interest* for the nod, as it can also be considered correct, but for simplicity's sake, we will first only consider *agree* as the correct answer). For the lean, *interest* will be considered the correct answer and for the head shake it is *disagree*. If we take another look at Table 4.1 we can see the number of participants that chose different answer options for each cue, the answers that will be considered as correct for now are highlighted in green.

Cue	Agree	Disagree	Interest	Disinterest	Other	Recognised	Answers
ChoreNod	36	0	6	0	1	38	43
ChoreNodEx	34	0	7	0	1	38	42
LiuNod	31	1	11	0	3	37	46
LiuNodEx	23	1	15	2	4	30	45
ChoreLean	1	0	31	2	7	39	41
ChoreLeanEx	2	0	26	3	11	36	42
FromScratchLean	2	0	29	2	8	38	41
FromScratchLeanEx	1	0	29	0	11	39	41
ChoreShake	1	35	2	4	1	38	43
ChoreShakeEx	0	33	4	8	2	38	47
FromScratchShake	0	32	4	9	1	39	46
FromScratchShakeEx	0	34	2	5	1	39	42

Table 4.1: Results from the validation online survey. The green regions show which of the answers was expected for this cue. The blue region shows another answer option which was considered correct for these cues and the yellow region highlights the highest numbers for the *Other* answer option. The cues that were chosen for the evaluation are printed in bold.

A chi-square test can be used to find out whether there is a statistically significant difference between expected and observed frequencies [Pan16]. The null hypothesis is that there is no significant difference. An observed frequency, in our case, is, for example, the number of participants that chose the *agree* option for the nod from Choregraphe. An expected frequency is the number of participants we expect to chose this option. So

the difference is that the observed frequency is an actual value, which can be looked up in Table 4.1 and the expected frequency is a theoretical value. This theoretical value will be calculated by weighting the correct and incorrect number of answers with the total number of answers given (done for each cue). Now we can check if our number of correct and incorrect answers is the same or close to our expected number of answers (expected frequency) by using the chi-square test.

Before the chi-square test can be conducted, a crosstab needs to be created with the data. The general structure of a 2 x 2 crosstab can be found in Figure 4.1. A more detailed explanation of crosstabs can be found on the website of Kent State University [Uni].

	Column 1	**Column 2**	**Row totals**
Row 1	a	b	$a+b$
Row 2	c	d	$c+d$
Column totals	$a+c$	$b+d$	$a+b+c+d$

Figure 4.1: A depiction of the structure of a 2 x 2 crosstab, taken from Kent State University's website [Uni].

To make this clearer, let us use our data and create our own 2 x 2 crosstab. We have two different conditions, a non-exaggerated cue (in this case the ChoreNod) and an exaggerated cue (ChoreNodEx), for example, and two possible outcomes: either the correct answer or an incorrect answer was selected. Now we can count the number of correct (1) and incorrect (0) answers for each cue and write it all down in a crosstabs (or square) in Table 4.2. These are the observed frequencies. Additionally, we add the sum of each row and column to the right and bottom of our square (see Table 4.2).

	0	1	$\sum$
ChoreNod - Observed	7	36	43
ChoreNodEx - Observed	8	34	42
$\sum$ - Observed	15	70	85

Table 4.2: Example for illustration of chi-square test using results from the validation survey for the nod from Choregraphe (ChoreNod and ChoreNodEx).

Now we have a crosstab of observed frequencies (the number of answers given was taken from validation survey results). The next step is to calculate the expected frequencies and add them to our table. For this, we first need the general structure of a 2 x 2 crosstab, including expected frequencies, which can be found in Figure 4.2. The final step

of data preparation is to add these calculations to our own table by using one of SPSS' *descriptive statistics* analysis features, the so-called *crosstabs* feature. We now need to input our variables into the rows and columns of the crosstabs, as depicted in Table 4.2. In the *cell display* window of SPSS' *crosstabs* feature, we can select that observed and expected values should be displayed (by ticking the box "expected"), which will result in SPSS displaying the expected values in the crosstabs. The output of the final crosstabs can be found in Figure 4.3.

	Column 1	**Column 2**	**Row totals**
Row 1	*a*	*b*	*a + b*
% of total	*a / (a + b + c + d)*	*b / (a + b + c + d)*	*(a + b) / (a + b + c + d)*
Row 2	*c*	*d*	*c + d*
% of total	*c / (a + b + c + d)*	*d / (a + b + c + d)*	*(c + d) / (a + b + c + d)*
Column totals	*a + c*	*b + d*	*a + b + c + d*
% of total	*(a + c) / (a + b + c + d)*	*(b + d) / (a + b + c + d)*	*(a + b + c + d) / (a + b + c + d) = 100%*

Figure 4.2: A depiction of the structure of a 2 x 2 crosstab, including expected frequencies (% of total), taken from the website of Kent State University [Uni].

Exaggerated * Agree Crosstabulation

			Agree		
			0	1	Total
Exaggerated	Non-exaggerated	Count	7	36	43
		Expected Count	7.6	35.4	43.0
	Exaggerated	Count	8	34	42
		Expected Count	7.4	34.6	42.0
Total		Count	15	70	85
		Expected Count	15.0	70.0	85.0

Figure 4.3: The results from the crosstab, as output by SPSS, for the nod from Choregraphe (ChoreNod and ChoreNodEx).

Now that we have prepared our data, we can conduct the chi-square test in SPSS by using

the *crosstabs* feature again. For this, we select *chi-square* as our *statistics* within the *crosstabs* feature. The results for the chi-square of the nod from Choregraphe (ChoreNod and ChoreNodEx), as output by SPSS, can be found in Figure 4.4. The exact equations used by SPSS to calculate the chi-square test can be found in IBM's manual for SPSS algorithms [IBM14, pages 151–152].

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.112[a]	1	.738		
Continuity Correction[b]	.003	1	.960		
Likelihood Ratio	.112	1	.738		
Fisher's Exact Test				.782	.480
Linear-by-Linear Association	.111	1	.739		
N of Valid Cases	85				

Figure 4.4: The results from the chi-square test, as output by SPSS, for the nod from Choregraphe (ChoreNod and ChoreNodEx). The numbers marked in red are the results that were used for this study.

Before we discuss the results from Table 4.4, the abbreviation *df* needs to be explained: it stands for degrees of freedom (*df*). The calculation of the degrees of freedom depends on the type of analysis that is done, and is therefore complex to describe. For this thesis, we will be using the values that are calculated by SPSS for the different types of tests. The same will be true for the df_{error}, which we will encounter in Section 4.2. The exact formulas used by SPSS to calculate *df* and df_{error} can be found IBM's manual for SPSS algorithms [IBM14, pages 29, 151].

The results from a chi-square test are commonly reported as [otUni10]:

$$\chi^2(df, N) = \chi^2\text{-value}, p = p\text{-value}$$

All the numbers that are relevant for the reporting of the chi square test results have been marked in red in Figure 4.4. Now we can start to fill in the reporting formula with the numbers from the table. The results for the chi-square (χ^2) can be found in column one (*Value*). The *df* can be read out directly from Table 4.3 (column two). The N in the formula stands for the number of total cases, which in the case of the chi-square test for the nod (ChoreNod and ChoreNodEx) is 85. The value for N can be taken from Table 4.4, from the first column and the last row (*N of Valid Cases*). The p refers to the p-value, which can also be taken from the table: It can be found in the first row and third column (*Asymptotic Significance (2-sided)*). The p-value is the probability (in

percent) that our results happened by chance. The results from the chi-square test from Figure 4.4 can be reported as:

$$\chi^2(1, 85) = 0.112, p = 0.738$$

Now we can report the results for all of our chi-square tests. A chi-square test was done to calculate the difference between the observed and expected frequencies for the non-exaggerated versus the exaggerated cues. Additionally, a test was done to calculate the difference between observed and expected frequencies for the cues created from Choregraphe versus the ones created from scratch or after Liu et al. [LIIH13]. This was done for all the cues (nod, lean and head skake). The results for all the computed chi-square tests can be found in Table 4.3.

Cues	$\chi^2(df, \mathrm{N})$	p
Non-Exaggerated versus Exaggerated		
ChoreNod | ChoreNodEx	$\chi^2(1, 85) = 0.112$	0.738
LiuNod | LiuNodEx	$\chi^2(1, 91) = 2.499$	0.114
ChoreLean | ChoreLeanEx	$\chi^2(1, 84) = 2.207$	0.137
FromScratchLean | FromScratchLeanEx	$\chi^2(1, 82) = 0.000$	1.0
ChoreShake | ChoreShakeEx	$\chi^2(1, 80) = 2.691$	0.101
FromScratchShake | FromScratchShakeEx	$\chi^2(1, 88) = 1.518$	0.218
Choreographe versus Created		
ChoreNod | LiuNod	$\chi^2(1, 89) = 3.185$	0.74
ChoreNodEx | LiuNodEx	$\chi^2(1, 87) = 8.563$	0.06
ChoreLean | FromScratchLean	$\chi^2(1, 82) = 0.248$	0.618
ChoreLeanEx | FromScratchLeanEx	$\chi^2(1, 83) = 0.723$	0.395
ChoreShake | FromScratchShake	$\chi^2(1, 85) = 0.285$	0.593
ChoreShakeEx | FromScratchShakeEx	$\chi^2(1, 89) = 1.375$	0.241

Table 4.3: The results from the qui-square tests. The two cues that were used in each test can be found on the left.

In order to find out which of these values are statistically significant, we need to take a look at the significance level (α), which is the threshold that determines whether the p-values from the results are statistically significant or not. A significance level commonly used is $\alpha = 0.05$ [otUni10], which results in a confidence interval of 95%. This means that if we test something that is, in fact, not statistically significant, we have a 95% chance of coming to the correct conclusion (meaning that it is not statistically significant) and only a 5% change of getting a Type 1 error (false positive) [Oan18]. This means that if the p-values from the results are the same as or below the threshold of the significance level

($\alpha = 0.05$), the results will be considered statistically significant for the validation survey. As can be seen from p-values in the table, none are statistically significant. One might argue that the Bonferroni correction needs to be computed [BA95] here. However, as it is used to prevent false positives and we do not have any positive results, it is not necessary to compute the Bonferroni correction in this case. What the Bonferroni correction is and how it is applied when there are positive results, will be discussed in Section 4.2.

4.2 Perception of Random, Contextual and Exaggerated Non-Verbal Cues

In this section, the quantitative results from the evaluation survey are discussed. The section is separated into the two different video types, the video where the robot was talking and the video where the robot was listening.

4.2.1 Robot Talking

For the statistical analysis of the evaluation survey data for the videos where the robot was talking, a repeated measures ANOVA was used. A repeated measures ANOVA is used when mean scores are measured under three or more different conditions. As the evaluation online survey measured the mean scores of three different conditions 1) random gestures, 2) contextual non-exaggerated gestures and 3) contextual exaggerated gestures, a repeated measures ANOVA is an appropriate statistical analysis.

An ANOVA is a statistical analysis which tests whether the mean scores of different groups are statistically significantly different from each other [Oan18]. Repeated measures means that the means of three or more different conditions are measured. We use this analysis to test whether the mean scores for our three conditions are statistically significantly different from each other.

To conduct the repeated measures ANOVA in SPSS, the analysis feature *repeated measures* from the *general linear models* was selected. Firstly, as this is a within-subject study, the within-subject variables and their levels need to be selected. We have five different within-subject variables, one for each Godspeed item (anthropomorphism, animacy, likeability, perceived intelligence and perceived safety) and the level for each one is three, as we have three different conditions (random gestures, contextual non-exaggerated gestures and contextual exaggerated gestures). Within-subject variables are independent variables where each participant (subject) has been exposed to each condition, in contrast to a between-subject variable, where each participant would be exposed to only one of the three conditions. Between-subject variables will be discussed in more detail at a later point in this section.

Before the repeated measures ANOVA can be conducted, however, the within-subject variables (Godspeed items) need to be created, as each one consists of the mean of the scores of all the questions that are a part of this Godspeed item. For example, for anthropomorphism the mean of the scores from the five questions (semantic differential scales):

A1 (*Fake* / *Natural*), A2 (*Machinelike* / *Humanlike*), A3 (*Unconscious* / *Conscious*), A4 (*Artificial* / *Lifelike*) and A5 (*Moving rigidly* / *Moving elegantly*) are used. This is done by using a transformation feature of SPSS, the *compute variables* feature. The code used for this in SPSS is the following (the A stands for anthropomorphism):

A = MEAN(A1, A2, A3, A4, A5)

With this code, the within-subject variable for anthropomorphism is created by calculating the mean of the scores for all the individual questions of this Godspeed item. The mathematical equation used by SPSS for the calculation of the mean can be found in IBM's manual for SPSS algorithms [IBM14, page 123].

With this method, a within-subject variable is created for each Godspeed item. As mentioned before, there is one within-subject variable for each Godspeed item and each one has a level of three (because of the three conditions). For example, we have the within-subject variable anthropomorphism, which has three levels, one with the mean score from the random condition, one with the mean score from the non-exaggerated contextual condition and one with the mean score from the exaggerated contextual condition. Therefore, the calculation of mean scores must be done three times for every Godspeed item (once with the scores from each level), which creates the three levels of the respective within-subject variable.

Now we can conduct our repeated measures ANOVA, but as a repeated measures ANOVA is only done for one independent variable at a time, we need to conduct five of them, one for each independent variable (Godspeed item). The mathematical equation SPSS uses for the calculation of the repeated measures ANOVA can be found in IBM's manual for SPSS algorithms [IBM14, page 490-493].

Before we take a look at the results, we need to make sure that all assumptions for a repeated measures ANOVA are met by our data. The assumptions for a repeated measures ANOVA are: 1) normality and 2) sphericity [Ltd]. Assumption 1) can be checked, for example, with a Shapiro-Wilk test [Oan18]. But as Ghasemi and Zahediasl [GZ12, page 486] state: "in large samples ($>$ 30 or 40), the sampling distribution tends to be normal, regardless of the shape of the data". Therefore, the Shapiro-Wilk test can be omitted. Assumption 2) can be checked by conducting *Mauchly's test of sphericity*. Mauchly's test of sphericity ensures that there is no violation of sphericity. That means that the variances of the differences between all possible combinations of the within-subject variables (Godspeed items) are equal. The mathematical equation used by SPSS for the calculation of the Mauchly's test can be found in IBM's manual for SPSS algorithms [IBM14, page 493-494]. An example table that contains the results of the tests, as output by SPSS, can be found in Figure 4.5.

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Anthropomorphism	.978	1.149	2	.563	.979	1.000	.500

Figure 4.5: Example of a results table from SPSS. The figure represents the results from Mauchly's test for the repeated measures ANOVA of the within-subject variable anthropomorphism. The numbers marked in red are the results that were used for this study. Sig. is short for significant and represents the p-value.

The results for Mauchly's test are commonly reported as follows [Ltd]:

$$\chi^2(df) = \chi^2\text{-value}, \; p = p\text{-value}$$

The symbols χ^2, df and p have already been explained in Section 4.1. The results from Figure 4.5 can be read as follows:

$$\chi^2(2) = 1.149, p = 0.563$$

The results for Mauchly's test for each Godspeed item can be found in Table 4.4. As the df are the same for every Mauchly's test that we conducted, it was put into the table header with a value of 2.

Godspeed Item	$\chi^2(2)$	p
Anthropomorphism	1.15	0.56
Animacy	0.72	0.7
Likeability	1.11	0.57
Perceived Intelligence	3.26	0.2
Perceived Safety	0.64	0.73

Table 4.4: The results for Mauchly's test for each Godspeed item. For each test result df is equal to 2.

As has already been explained in Section 4.1, the p-values are significant if they are on or below the significance level α, which was set at 0.05. When looking at the results from Mauchly's test in Table 4.4, we can see that none of the p-values are below the threshold of 0.05, which means that no violation of sphericity could be found and we can continue by looking at the results of our repeated measures ANOVA, in Figure 4.6.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Anthropomorphism	Sphericity Assumed	3.861	2	1.931	4.814	.010
	Greenhouse-Geisser	3.861	1.957	1.973	4.814	.010
	Huynh-Feldt	3.861	2.000	1.931	4.814	.010
	Lower-bound	3.861	1.000	3.861	4.814	.033
Error(Anthropomorphism)	Sphericity Assumed	42.512	106	.401		
	Greenhouse-Geisser	42.512	103.734	.410		
	Huynh-Feldt	42.512	106.000	.401		
	Lower-bound	42.512	53.000	.802		

Figure 4.6: Example of a results table from SPSS. The figure depicts the results from the repeated measures ANOVA for the within-subject variable (or within-subject effect) anthropomorphism. The numbers marked in red are the results that were used for this study. F represents the F-value.

When looking at the results from Figure 4.6, we can see that some results are marked red. These are the results that will be used in this study. The results for a repeated measures ANOVA are commonly reported as follows [otUni10]:

$$F(df, df_{error}) = \text{F-value}, p = p\text{-value}$$

In Section 4.1, we have already talked about df, df_{error} and the p-value, but not about the F-value. While the p-value represents a probability of getting incorrect results, the F-value is a test statistic, which is a ratio of variances of our observations [To].

The results for the repeated measures ANOVA for anthropomorphism can now be put into the formula for reporting. The values are marked red in Figure 4.6. The results for the repeated measures ANOVA for anthropomorphism can be reported as:

$$\text{F}(2,\ 106) = 4.81,\ p = 0.010$$

The results from the other repeated measures ANOVAs (for each of the Godspeed items) were calculated and read from the results table in the same way and can be found in Table 4.5.

As mentioned before, in order to find out which of these values are significant, we need to take a look at the significance level (α). As we have conducted five ANOVAs with the same data, the problem of multiple comparisons needs to be addressed. The problem of multiple comparisons refers to the fact that if multiple hypotheses are tested on one data set, the probability of encountering a rare event increases, which in turn means the

probability of making a Type 1 error increases. A Type 1 error occurs when the null hypothesis is rejected incorrectly, meaning the results show a false positive. In order to address this, a Bonferroni correction can be done [BA95], where the significance level α is adjusted to α'.

Godspeed Item	F(2, 106)	p
Anthropomorphism	4.814	0.010*
Animacy	9.119	0.000**
Likeability	3.066	0.051
Perceived Intelligence	2.95	0.057
Perceived Safety	6.787	0.002*

Table 4.5: The results for the repeated measures ANOVA for each Godspeed item. For each F-value in this table df is equal to 2 and df_{error} is equal to 106, therefore these values have been put into the column header. Statistical significance from the repeated measures ANOVA is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

For the Bonferroni correction, the following calculation was used in our study [BA95]:

$$\alpha' = \alpha/n$$

α' is the new α, which is calculated by dividing α by the number of times the analysis was conducted with the same data set. As we conducted the repeated measures ANOVA five times (once for each Godspeed item), our n is equal to five. Therefore the calculation was done as follows:

$$\alpha' = 0.05/5 = 0.01$$

The new significance level α' is therefore 0.01. Meaning that the p-values that are the same as or below this threshold are statistically significant. Now we can look at the p-values (from Table 4.5) and find out whether they are statistically significant. The p-values for anthropomorphism ($p = 0.010$), animacy ($p = 0.000$) and perceived safety ($p = 0.002$) are the same as or below the threshold of α' (α' $= 0.01$) and therefore statistically significant. As can be seen in Table 4.5 the p-values that are statistically significant are marked with one asterisk (where $p \leq 0.01$) or two asterisks (where $p \leq 0.001$), this will be the notation used for the rest of this thesis.

The mean score of the individual categories of the Godspeed questionnaire and the different types of cues for the robot-talking condition can be found in Figure 4.7. The Godspeed items for which the repeated measures ANOVA provided statistically significant results are marked with asterisks.

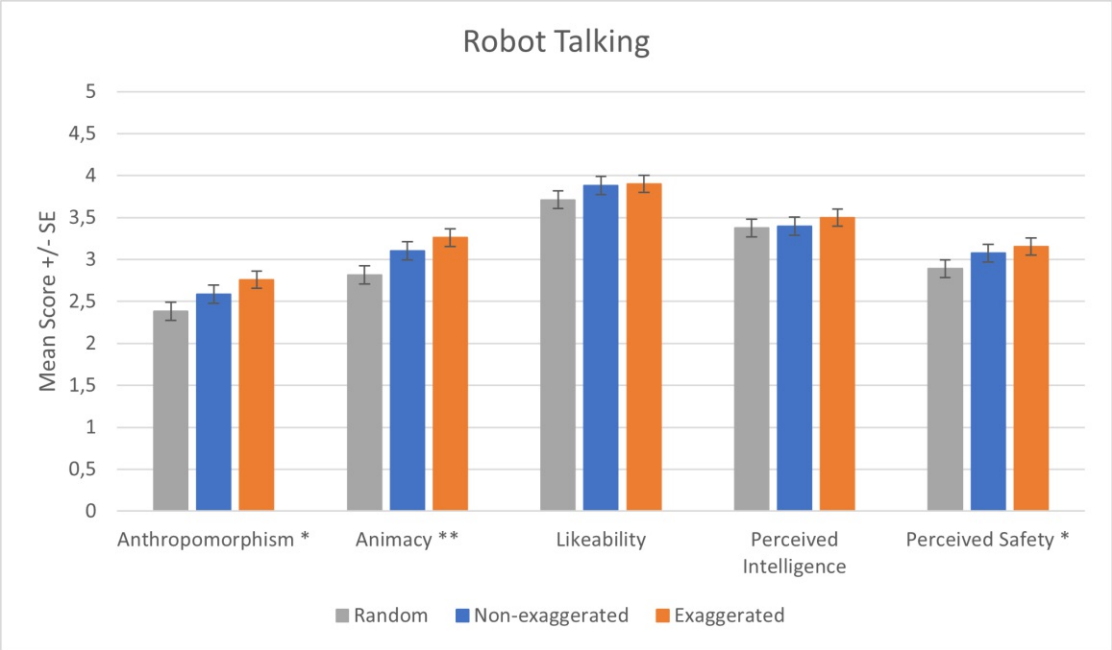


Figure 4.7: A bar chart describing the results of the evaluation survey for the robot-talking condition, for each Godspeed item. Statistical significance from the repeated measures ANOVA is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

For the repeated measures ANOVA, five between-subject variables were checked for statistical significance: the participant's gender, age, level of education, experience with robots and whether or not they were a student. The country of origin was not included, because only four (7.41%) participants did not report *Austria* as their country of origin.

Between-subject variables (or factors) are (in contrast to within-subject factors) measured only once and therefore, we always compare the results between different groups of participants. An example for a between subject variable is gender, as every participant can only be in one of the gender groups (male, female, non-binary or did not disclose).

The between subject factors can be specified in SPSS when a repeated measures ANOVA is conducted, by adding the variables (such as gender, age, ect.) to the list of *between-subject factor(s)* in the SPSS interface of the repeated measures ANOVA. However, as none of these between-subject effects were statistically significant they were not included in the thesis. The significance was determined the same way it was done for the other statistical tests in this section. The relevant column of the results was the *Sig.* column that shows the p-value for every between-subject effect. However, none of these values were below the threshold of α and therefore not statistically significant.

Since each Godspeed item consists of three to six questions (semantic differential scales), a second repeated measures ANOVA was done with the individual questions in order to find out which individual questions are statistically significant. For this, no calculation of

within-subject variables is necessary because instead of using the mean of the individual questions of each Godspeed item, we are directly using the scores for each individual question. The Bonferroni correction had to be done individually for each Godspeed item because they do not all have the same number of questions. Table 4.6 shows the results for the Mauchly's tests and the repeated measures ANOVA for the individual questions concerning animacy. The calculation of the test and the repeated measures ANOVA were done in the same way as they were done previously in this section. Therefore, for the repeated measures ANOVA of the individual questions, only the results for animacy are presented. The table for all the other individual questions can be found in Appendix B. The table also shows the new α' calculated with the Bonferroni correction and whether the results were significant (marked with asterisks). As indicated in the table, the results were significant for animacy questions one (*Dead* / *Alive*), two (*Stagnant* / *Lively*) and six (*Apathetic* / *Responsive*). All the other Godspeed items did not have any statistically significant results for the individual questions.

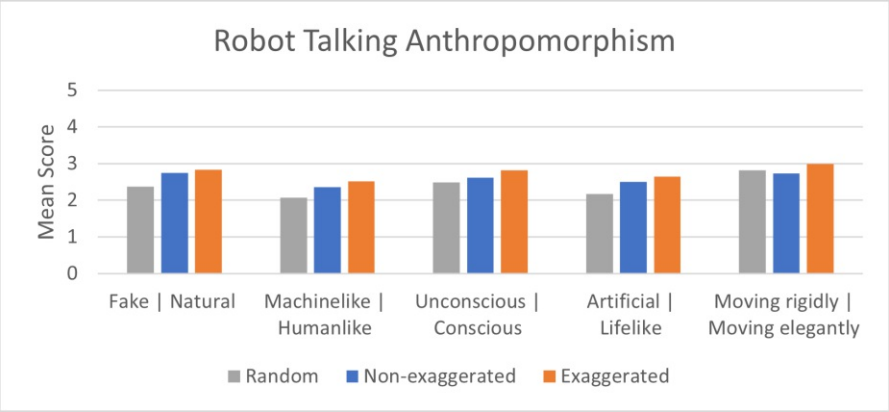
Semantic Scale	Mauchly's $X^2(2)$	Mauchly's p	ANOVA F(2, 106)	ANOVA p	α'
Animacy					
Dead/Alive	5.666	0.056	7.179	0.001**	0.008
Stagnant/Lively	1.391	0.499	7.178	0.001**	0.008
Mechanical/Organic	1.649	0.439	3.183	0.045	0.008
Artificial/Lifelike	1.573	0.455	4.74	0.011	0.008
Inert/Interactive	2.205	0.332	2.886	0.6	0.008
Apathetic/Responsive	3.761	0.153	6.638	0.002*	0.008

Table 4.6: Results from ANOVA done for the individual questions of each of the Godspeed items of the robot-talking condition. Statistically significance is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

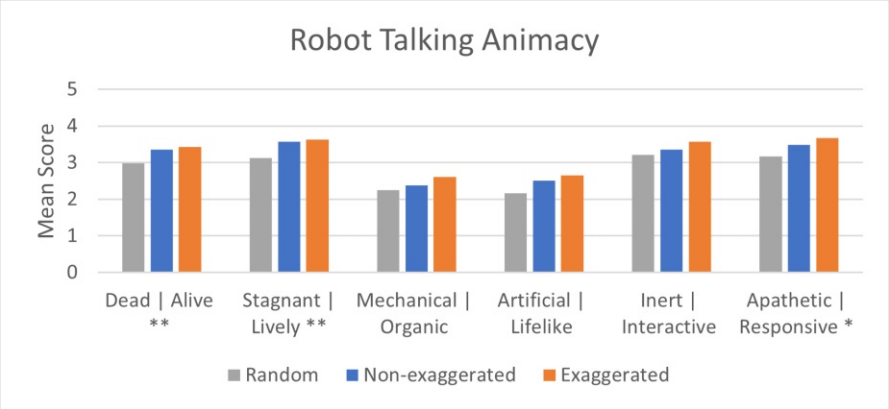
The mean score for each individual question of each of the significant Godspeed items for the robot-talking condition can be found in Figure 4.8. The bar charts containing the individual questions from perceived intelligence and perceived safety can be found in Appendix B. When talking about the results from the individual questions of each of the Godspeed items, the results from the ANOVA over the Godspeed items (not the individual questions of each item) will be referred to as combined Godspeed item results, as they contain the results for the individual questions combined. An example would be the combined anthropomorphism results.

When examining Figure 4.8a, it can be observed that the bars of all the questions look similar to those from the combined anthropomorphism results, as the random condition is mostly rated lowest and the exaggerated condition is mostly rated highest. The only exception is the last question (*Moving rigidly* / *Moving elegantly*), where the random condition is rated higher than the non-exaggerated condition.

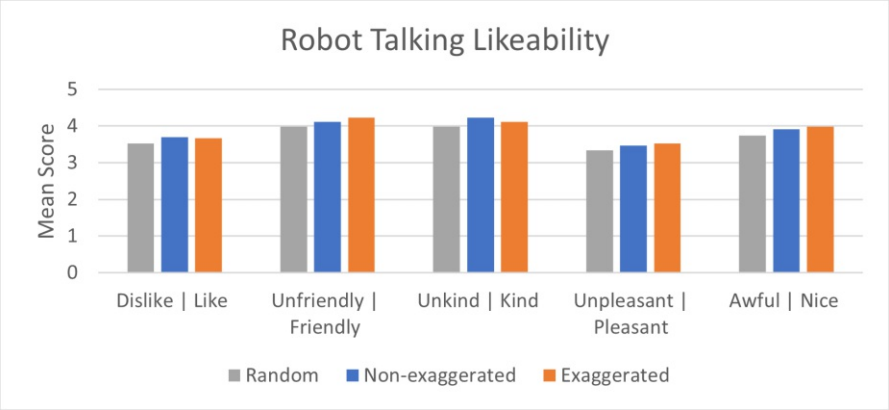
As mentioned previously, for animacy (see Figure 4.8b) three of the six questions were statistically significant: the first question (*Dead* / *Alive*), the second question (*Stagnant* / *Lively*) and the sixth question (*Apathetic* / *Responsive*).



(a) The results for the individual questions of the item anthropomorphism.



(b) The results for the individual questions of the item animacy.



(c) The results for the individual questions of the item likeability.

Figure 4.8: Three charts describing the results for the individual questions of the Godspeed items, for the robot-talking condition. The y-axes represent the mean values of the Godspeed score and the x-axes show the different questions (semantic differential scales). Statistical significance is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

The bar charts for likeability (see Figure 4.8c) show that the results for questions four (*Unpleasant* / *Pleasant*) and five (*Awful* / *Nice*) resemble the results from the combined likeability score. However, for question one (*Like* / *Dislike*), the results are different as the non-exaggerated condition was rated highest and the exaggerated condition in the middle. For the questions two (*Unfriendly* / *Friendly*) and three (*Unkind* / *Kind*) all the different conditions were rated the same and achieved the highest rating possible.

4.2.2 Robot Listening

For the statistical analysis of the data from the evaluation survey for the videos where the robot was listening, a paired t-test was conducted. A paired t-test (also paired samples t-test) is used to find out whether the mean scores of two groups are statistically significantly different from each other, similar to the repeated measures ANOVA, but with only two conditions (in this case non-exaggerated cues and exaggerated cues) instead of three or more [Oan18]. The program SPSS was used for all the calculations.

As a first step for the paired t-test the five within-subject variables (one for each Godspeed item) need to be calculated. This was done the same way as it was done for the repeated measures ANOVA, as described in Section 4.2.1. Only this time, the data of the videos where the robot was listening, instead of the ones where the robot was talking, was used.

For a paired t-test, instead of testing for sphericity when there is only a small number of participants (below approx. 30 [SW65]) a test for normal distribution is often done (such as the Shapiro-Wilks test) [Oan18]. However, as the number of participants in this study was 54 there is no need to run this test before calculating the paired t-test.

For the paired t-test, SPSS' *paired samples t-test* from the analysis feature *compare means* was used. As a first step, pairs need to be selected. A pair consists of the results of two conditions or levels (the non-exaggerated condition and the exaggerated condition) of a within-subject variable (one for each Godspeed item). For example, for anthromophism, the results of the non-exaggerated condition are one part of the pair (level 1) and the results of the exaggerated condition are the other part of the pair (level 2). They are a pair, because they measure the same scores, only in a different condition. The formulas used by SPSS for the calculation of the paired t-test can be found in IBM's manual for SPSS algorithms [IBM14, pages 907-908]. An example for results as output by SPSS can be found in Figure 4.9.

The results from the paired t-test are commonly reported as follows [otUni10]:

$$t(df) = t - value, p = p - value$$

Additionally, the values for mean (M) and the standard deviation (SD) of each level of

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	RLNA	2.1259	54	.77440	.10538
	RLEA	2.3000	54	.88914	.12100

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	RLNA - RLEA	-.17407	.73720	.10032	-.37529	.02714	-1.735	53	.089

Figure 4.9: Example of a results table from SPSS. The figure depicts the results from the paired samples t-test for anthropomorphism. RLNA stands for the non-exaggerated condition and RLEA stands for the exaggerated condition. The numbers marked in red are the results that were used for this study. The *df* stands for degrees of freedom, t is the t-value and Sig. (2-tailed) represents the *p*-value.

the within-subject variable are commonly reported. The t-value is described by Oancea [Oan18] as:

> "Each t-test calculates a test statistic (called t-value in this case). The t-value (or t-score) represents the ratio between the difference of the group means and the differences within the groups." [Oan18, page 67]

When looking at the results for anthropomorphism from Figure 4.9, the results can be reported as:

$$\text{Non-exaggerated (M = 2.126, SD = 0.774), exaggerated (M = 2.3, SD = 0.889),}$$
$$\text{t(53) = -1.741, } p = 0.089$$

Godspeed Item	Non-Exag.(M)	Non-Exag.(SD)	Exag.(M)	Exag.(SD)	t(53)	p
Anthropomorphism	2.126	0.774	2.3	0.889	-1.735	0.089
Animacy	2.679	0.765	2.958	0.883	-3.008	0.004*
Likeability	3.437	0.784	3.622	0.82	-1.635	0.108
Perceived Intelligence	3.133	0.57	3.2	0.643	-0.956	0.343
Perceived Safety	2.895	0.582	3.34	0.684	-4.529	0.000**

Table 4.7: The results for the paired t-test for each Godspeed item. For each t-value in this table, *df* is equal to 53, therefore this value has been put into the column header. Statistical significance from the paired t-test is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

Just as with the repeated measures ANOVA, a separate test was carried out for each Godspeed item. Therefore, the Bonferroni correction needs to be applied for the paired t-test as well. That means that the significant value $\alpha' = 0.01$. This means that a result is statistically significant when p is on or below the threshold of $\alpha' = 0.01$. For a more detailed explanation, see Section 4.1.

Therefore, the results are statistically significant for animacy ($p = 0.004$) and perceived safety ($p = 0.000$) for the videos where the robot was listening.

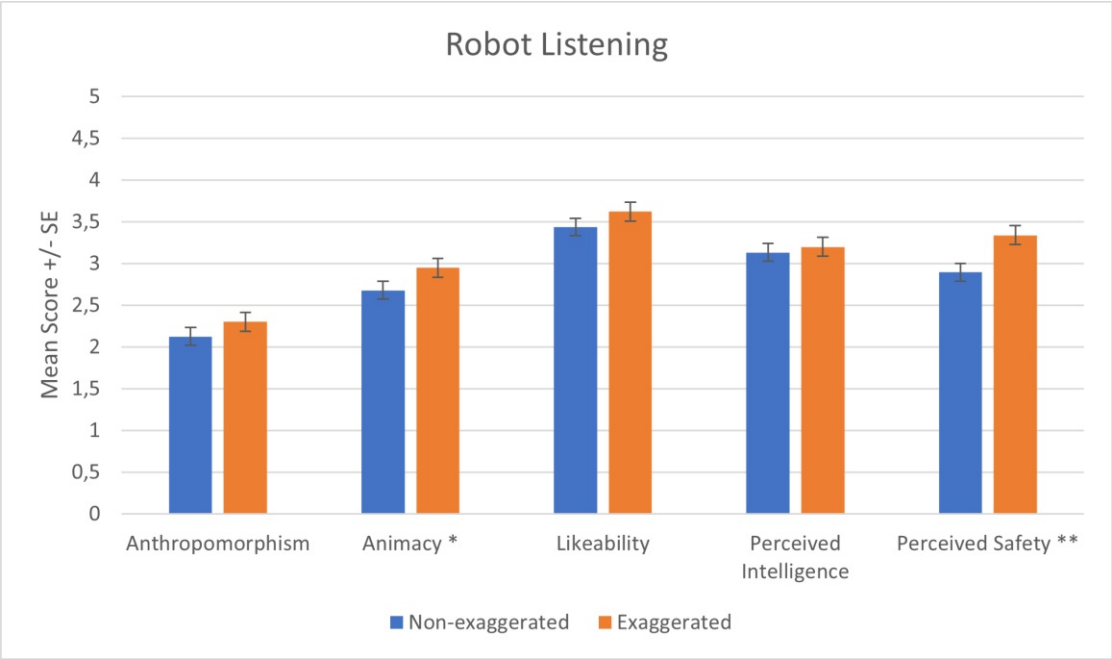


Figure 4.10: A bar chart describing the results of the evaluation survey for the videos where the robot was listening. Statistical significance is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

The mean score of the Godspeed items and different types of cues for the videos where the robot was listening can be found in Figure 4.10. The items for which the results were statistically significant are animacy (the second set of bars) and perceived safety (the last set of bars), which are marked with asterisks.

Just as in Section 4.2.1, additional t-tests were done for the individual questions (semantic differential scales) of each Godspeed item for the videos where the robot was listening. The t-test was conducted with the same methods as described earlier in this section. The process of analysing the individual Godspeed questions has already been explained in detail in Section 4.2.1. The same methods were used, the only difference being that a paired t-test was used instead of a repeated measures ANOVA.

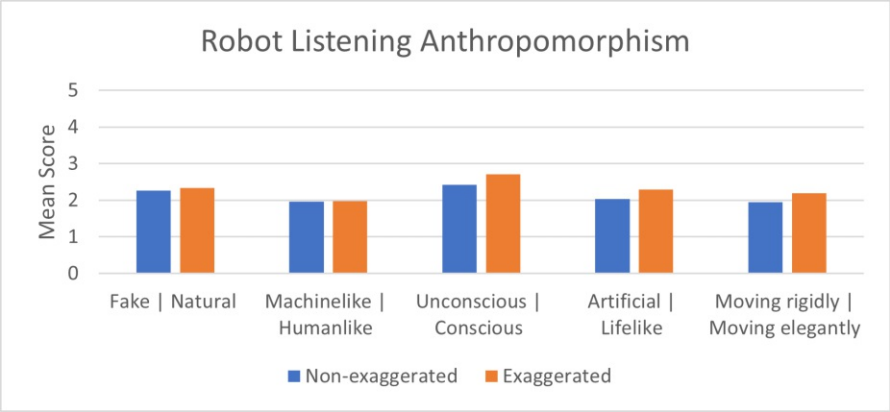
The results for each of the individual questions (semantic differential scales) of each Godspeed item for the videos where the robot was listening can be found in Appendix B. Two sections of the table containing the results can be found in Table 4.8. These sections (the results for animacy and perceived safety) were selected as they contain all the statistically significant results. The results are statistically significant for the semantic differential scales *Stagnant / Lively* and *Apathetic / Responsive* from animacy and *Calm / Agitated* and *Quiescent / Surprised* from perceived safety (marked by asterisks in the table).

Semantic Scale	Non-Exag.(M)	Non-Exag.(SD)	Exag.(M)	Exag.(SD)	t(53)	p	α'
Animacy							
Dead/Alive	2.85	0.960	3.02	1.141	-1.456	0.151	0.0083
Stagnant/Lively	2.89	1.093	3.31	1.043	-2.961	0.005*	0.0083
Mechanical/Organic	1.94	0.940	2.02	0.942	-0.562	0.576	0.0083
Artificial/Lifelike	2.04	0.910	2.30	1.127	-1.847	0.070	0.0083
Inert/Interactive	3.07	1.025	3.31	1.195	-1.788	0.079	0.0083
Apathetic/Responsive	3.28	0.856	3.72	0.940	-3.385	0.001**	0.0083
Perceived Safety							
Anxious/Relaxed	3.43	1.021	3.48	1.005	-0.454	0.652	0.0167
Calm/Agitated	2.37	0.996	3.15	1.219	-5.071	0.000**	0.0167
Quiescent/Surprised	2.89	1.058	3.39	0.998	-2.860	0.006*	0.0167

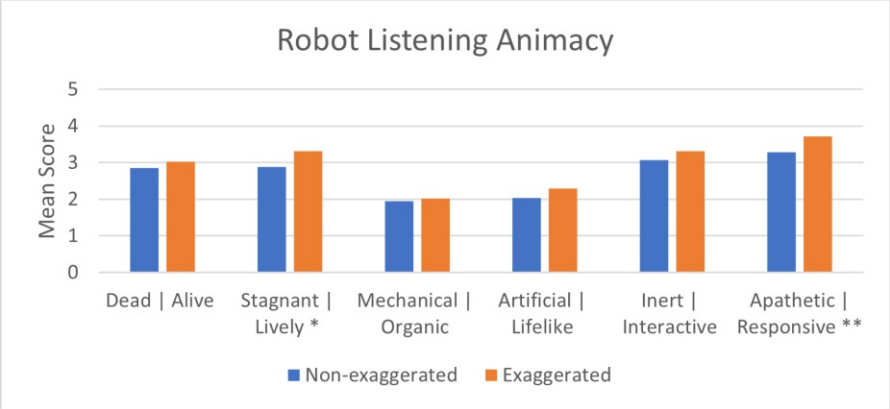
Table 4.8: Results from paired t-test done for the individual questions of each of the Godspeed items of the robot-listening condition. Statistical significance is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

The mean score for the individual questions for the robot-listening condition can be found in Figure 4.11. Just as the figure for the robot-talking condition this figure shows three bar charts, as two of the charts for the individual Godspeed questions do not contain information used in this thesis. The omitted charts for likeability and perceived intelligence can be found in Appendix B. The statistically significant values are marked with asterisks and can be found in Figure 4.11b and Figure 4.11c.

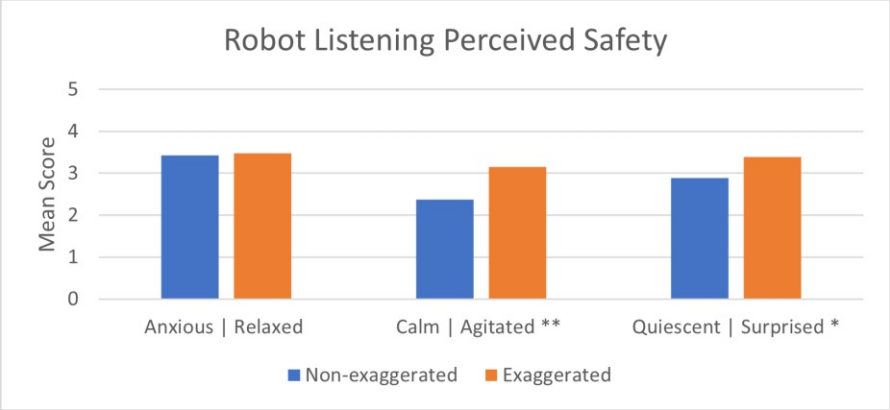
For anthropomorphism (see Figure 4.11a), the bars are similar to the ones from the combined anthropomorphism results. For question two (*Machinelike / Humanlike*) the bars are almost the same height, meaning the difference between the two conditions was very small. For animacy (see Figure 4.11b), all the questions are again similar to the collective results, but questions two (*Stagnant / Lively*) and six (*Apathetic / Responsive*) stand out as there are much higher differences between the two conditions than for the other questions. In the chart for likeability, question four (*Unpleasant / Pleasant*) shows the opposite result of all the other questions, with non-exaggerated being rated slightly higher than exaggerated. Lastly, for perceived safety (see Figure 4.11c), question one (*Anxious / Relaxed*) showed almost no difference between the two conditions while question two (*Agitated / Calm*) showed a very clear difference.



(a) The results for the individual questions of the item anthropomorphism.



(b) The results for the individual questions of the item animacy.



(c) The results for the individual questions of the item perceived safety.

Figure 4.11: Three charts describing the results for the individual questions of the Godspeed item, for the robot-listening condition. The y-axes represent the mean values of the Godspeed score and the x-axes show the different questions (semantic differential scales). Statistical significance is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

4.3 People's Experience of Different Non-Verbal Cues

The data from the open-ended questions from the evaluation online survey and the data collected from the four semi-structured interviews was analysed via thematic analysis. The aim is to gain more insight into people's experiences and concerns and to find common themes in their experiences. The themes and sub-themes have been visualised using a sunburst diagram, which can be seen in Figure 4.12. The major themes that emerged during the analysis are listed and explained in more detail here:

Likeable Robot

Regarding the robot in general, all interviewees described the robot with positive adjectives such as friendly, likeable, cute, interested and understanding. The overall theme is that the robot makes a positive impression. Concerning the robot's movement, the overarching theme was mechanical, with the words stiff and jerky being used to describe it. The interaction of the robot was described as a little bit unnatural but in general working well. Multiple people stated that the timing was essential and that it looked very convincing and well done in the videos. It was also stated by some that the robot positively surprised them with its interaction skills.

Relaxed but Stiff Situation

The situation was described as relaxed by most people. Some described it as a little bit stiff but this was due to the robot's stiffness. They described the setting as a bit unnatural which created an atmosphere that reminded them of an experiment or a test. The interaction was described as natural, despite the setting.

Movement over Stillness

There was also the theme of movement being highly valued: not only were people positively surprised by how much the robot moved, but they also stated that even more movement would be even better. The moments in the videos where the robot did not move were reported as appearing mechanical and machine-like.

Exaggeration Preferred

Although the overall theme confirmed that exaggeration was preferred, it was also described as less human-like but cuter and more child-like. The non-exaggerated movements were described as more subtle as well as more natural and less exaggerated in appearance. Nonetheless, the exaggerated versions were rated more favourably overall. Some people mentioned that the difference between exaggerated and non-exaggerated was hardly noticeable, especially when the videos were not immediately followed by one another. The biggest difference they reported was between the non-exaggerated and exaggerated head shaking in the robot-listening condition.

Figure 4.12: A sunburst diagram visualising the themes and sub-themes of the thematic analysis of the interviews and open-ended survey questions. The colours represent the themes, starting with super-themes on the inside circle, themes in the middle and sub-themes on the outside circle. The size represents the frequency of appearance.

Contextual over Random

For the robot-talking condition, another theme that emerged was that contextual movement was overall preferred to random movement. The random movement was described as more human-like, but the contextual movement, especially the exaggerated movement, made the intent of the robot easier to interpret. Some participants suggested a mix of

random and contextual movement.

Cracking Sound

During some movements, especially the forward lean and even more so during the exaggerated forward lean, a loud cracking sound, originating from the hip joint, could be heard. There was also sounds coming from the motors, which were especially audible when the robot was shaking its head exaggeratedly. People reported this to be very distracting and suggested that it made the robot appear more mechanical and less likeable. The biggest problem related to these sounds was reportedly how distracting they were. This was less of a problem for the head shaking, which was generally still perceived well. The cracking of the hip, however, was reported as very unpleasant.

Annoying Voice

Another theme was the mechanical and high pitched voice of the robot, which reportedly lacked any kind of natural intonation and pauses. The voice was also described as sounding nasal and generally unpleasant. Some people, however, said it was what they expected of a robot and the voice fit the robot's appearance.

No Facial Expressions

A slightly smaller but still present theme was the robot's lack of facial expressions. Especially when the robot was talking, the complete absence of a mouth or any other facial movement reportedly made it feel much more like a machine than a person.

CHAPTER 5

Discussion

In this chapter, the three research questions that were posed at the beginning of this work (see Section 1.2) are answered. The first research question focuses on the readability of different non-verbal cues and to answer it, the results from the validation survey are analysed. The second research question is separated into the robot-talking condition and the robot-listening condition and focuses on the perception of non-verbal cues performed by Pepper. Lastly, the third research question is about the experience people had when watching Pepper perform different non-verbal cues. To answer these questions, the qualitative data from the open-ended survey questions and the interviews is analysed. The chapter finishes with the limitations of this study.

5.1 Research Question 1

How do different ways of creating non-verbal cues for the robot Pepper impact their readability?

When examining readability, it was found that all cues were recognised well and no statistically significant difference between the factor non-exaggerated versus exaggerated or the factor of the different creation methods, could be found. Therefore, Hypothesis 1 (see Section 1.2), which stated that the exaggerated cues would be more readable was not supported by the data. This could be related to the way the validation questions were phrased: a semantic scale, for example, might have allowed for more nuanced answers. As already mentioned in Chapter 3, Atkinson et al. [ADGY04] found that exaggeration increased the readability of emotions, except for sadness. They also used videos in their study, however, as explained previously, they used humans and light strips instead of robots. Conclusively it can be said that no significant difference in the readability of differently created non-verbal cues (exaggerated, non-exaggerated, from Choreographe or from scratch) could be found.

As the results of the validation survey were very similar for all the cues, it was decided by the authors that the nod created after Liu et al. [LIIH13] would be used in the evaluation. This was decided even though the results for matching the from-scratch nod with the word *Agree* were higher than the corresponding results for the nod created after Liu et al. [LIIH13]. The reason for this is that in the evaluation, the nod would be used during the robot-listening condition as a way for the robot to indicate that they were listening and understood the story. Therefore, the nod created after Liu et al. [LIIH13] was chosen because it had higher results for the word *Interest*, which was more in line with what the robot was supposed to convey during the evaluation. For the lean and the head shake it was decided, as all the ways of creating non-verbal cues had yielded similar results, that one created after Choregraphe and one created from scratch should be used. The head shake from Choregraphe and the lean created from scratch were chosen.

5.2 Research Question 2

How do random, contextual and exaggerated non-verbal cues influence a human's perception of anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of the robot Pepper?

5.2.1 Robot Talking

When examining how different non-verbal cues influence the perception of the robot Pepper, the focus lies firstly on exaggerated versus non-exaggerated cues (see Hypothesis 2.1 in Section 1.2) and secondly on random versus contextual cues (see Hypothesis 2.2 in Section 1.2).

When considering the results for the robot-talking condition, Hypothesis 2.1 was partially confirmed, as it states that exaggerated cues would be rated higher for all Godspeed items than non-exaggerated cues. The results were statistically significant for anthropomorphism, animacy and perceived safety. That means that exaggerated cues were generally perceived more favourably in these Godspeed items than non-exaggerated cues. Hypothesis 2.2 states that contextual cues would be rated higher than random cues. This was also partially confirmed and statistically significant for the same three Godspeed items. This means that contextual cues were perceived more favourably than random cues in these Godspeed items. Both of these results are also supported by the qualitative data, which suggests that exaggerated cues were perceived as clearer, cuter, funnier and more childish when compared to non-exaggerated and random cues.

Gielniak and Thomaz [GT12] found that exaggerated motion used by the SIMON robot yielded higher scores in engagement and entertainment. This statement is supported by the results from this study, especially because of the positive results for animacy, which represents the liveliness of the robot and determines as how interactive it is perceived.

The anthropomorphism being statistically significant (see Section 4.2) and the exaggerated cues receiving a higher rating than the non-exaggerated cues contradicts the findings from the interviews, where multiple people said that the exaggerated version was more

comical and less human-like. This is an interesting aspect and in order to find out more about it, additional research should be done, including more interviews and another survey with a larger and more diverse group of participants.

When looking at the individual questions related to animacy from the five Godspeed items, the results were only statistically significant for the following semantic scales: *Dead / Alive*, *Stagnant / Lively* and *Apathetic / Responsive*. Especially the last of the three supports the findings from Gielniak and Thomaz [GT12], who found that exaggeration increases the perception of engagement, which is similar to responsiveness.

The rest of the individual questions, though not statistically significant also hold some interesting findings. For anthropomorphism, the semantic scale *Moving rigidly / Moving elegantly* shows an unusually high rating for the random condition. This might be because the random movements were constant and, unlike the contextual movements, not only present when the robot was talking about something specific. Therefore, the movements probably appeared more fluent / elegant.

Another interesting set of individual questions to look at are the ones concerning likeability. There was an inconsistency with the combined score for *Dislike / Like*, which was that the combined score for likeability showed the highest scores for the exaggerated condition and the individual score for question one *Dislike / Like* showed the highest scores for the non-exaggerated condition. This might be related to the official German translation which was *Nicht mögen / Mögen*, might have confused people by making them think that they had to rate whether or not the robot liked or disliked something.

5.2.2 Robot Listening

For the robot-listening condition, we will only look at Hypothesis 2.1, as there were no random cues in this condition. Hypothesis 2.1 was again partially supported as the results were statistically significant for animacy and perceived safety. This again supports the findings by Gielniak and Thomaz [GT12], just as in the robot-talking condition.

For the individual questions concerning animacy, the *Stagnant / Lively* and *Apathetic / Responsive* were statistically significant. When it comes to perceived safety, the *Agitated / Calm* and *Quiescent / Surprised* scales were statistically significant. As the two questions from animacy have already been discussed in the robot-talking condition, we will focus on perceived safety. The results for two out of the three questions were statistically significant, which is not surprising, considering that the result for all questions about animacy combined was statistically significant as well. But what is interesting is the fact that for the robot-talking condition none of results for the individual questions from perceived safety were statistically significant. This might be due to the fact that while the robot was listening it was not moving as much and therefore received a higher rating for being calm and quiet. The exaggerated head shake (from the robot-listening condition), however, was more extreme than most of the movements from the robot-talking condition. More extreme meaning that the movement was strongly exaggerated

and repeated multiple times. Therefore, the differences in rating between non-exaggerated and exaggerated were larger than when the robot was talking.

When looking at the results for the individual questions which were not statistically significant, we see that some of them show very little difference between the non-exaggerated and the exaggerated condition. This might be due to the fact that the exaggerated version of the cues was not very extreme in the robot-listening condition, the head shake features being the most noticeable difference. This is supported by the qualitative data: some participants stated that they had a hard time noticing the difference between the cues from the two conditions (except for the head shake). It would be interesting to continue this research with more extreme exaggeration, as is might lead to larger differences between the two conditions. However, the fact that most of the questions still have a higher rating for the exaggerated cue and some of them are even statistically significant implies that people likely subconsciously noticed an aberration and that is what is reflected in the score they gave.

5.3 Research Question 3

How do people describe their experiences of watching different non-verbal cues being expressed by the robot Pepper in the context of a conversation?

Overall, people reported a pleasant experience and the robot was perceived favourably (see Figure 4.12). Some of the qualitative findings are supported by the quantitative findings from RQ2 and can therefore be found in Subsection 5.2.2 and 5.2.1. Especially interesting were the concerns people expressed about the robot. The biggest concern was the loud cracking noise that Pepper's hip joint made when the robot was leaning forward. This was reported as being distracting and unpleasant. This is a useful insight for future research, which an exclusively quantitative data collection would probably not have been able to provide. The movements that create such loud noises should either be avoided, if possible, or performed more gently, which might reduce the stress on the joints. This is also something that should be investigated in future research.

The data from our interviews which describe the robot using exaggerated cues as comic-like and animated confirm the results by Gielniak and Thomaz [GT12] who found that exaggeration made the robot appear more comic-like.

5.4 Limitations

Firstly, there was a bias in the people that participated in the study, as most of them were students of computer science around the age of 25 years old and almost all of them were from Austria. Additionally, due to the nature of conducting an anonymous online survey it could not be ensured that participants did not participate more than once, as no cookies were used or IP addresses saved, in order to ensure data protection.

Naturally, only using video instead of in-person interaction is also a limitation for this study. In-person studies were partly limited due to the Covid-19 pandemic.

Lastly, partly a limitation but mostly a lesson learned was that the way the validation survey questions were asked made it difficult to statistically analyse them. If the questions had been Likert or semantic scales (as they were for the evaluation survey), it would not only have given more opportunity for statistical analysis, but also more in-depth information, since a scale can provide more information than just the option to tick certain answers.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

The goal of this research was to answer questions about the creation, readability and perception of non-verbal cues for the robot Pepper. Different cues (nodding, leaning, head shaking and gestures) were created by applying three different methods: using pre-implemented cues from the program Choregraphe, using cues from literature [LIIH13] and using cues created from scratch with inspiration from literature. Two online surveys and four semi-structured interviews were conducted. Although no significant differences in readability could be found in any of the conditions, the results show that the robot was perceived more positively in the categories of animacy and perceived safety of the Godspeed questionnaire when it used exaggerated instead of non-exaggerated cues. Additionally, the hypothesis that stated that the robot would be perceived more negatively when using random cues in contrast to contextual cues was confirmed for these Godspeed items. For the condition where the robot was talking, related results were found to be statistically significant for anthropomorphism. The results from the qualitative data support the results from the quantitative data, as a thematic analysis of the open-ended questions from one of the online surveys and the interviews confirmed that exaggeration was described more positively than the other conditions. Our investigations also bring specific concerns to light, such as the loud and distracting nature of the sound that the robot's motors produce during certain movements.

6.2 Future Work

As this research used videos to find out how humans, perception of robots can be influenced by the type of non-verbal cues the robot is using, it would be useful to do similar tests with in-person interaction. This would show whether the results from the

videos will prove to be similar to the results from live interaction. Such findings would be very useful, as robots are intended for in-person interaction.

Additionally, it would be interesting to find out if there is a cap to the exaggeration being perceived as more favourable in some categories of the Godspeed questionnaire. If the exaggeration was much more extreme, even including more repetition, would there be a point at which the results would tip into the other direction? This would be another interesting path for future research.

To investigate the topic of this thesis further, a study with even more participants could be carried out. There could also be an additional mode to random, contextual and exaggerated, namely a mix of random and contextual movements, which was suggested by some of the interviewees.

The addition of breathing and other natural motions that are pre-implemented in the robot could also be a next step of testing exaggeration of non-verbal cues in a less isolated and more realistic fashion. This could also counteract the problem that a lack of movement made the robot appear somewhat mechanical, which was reported by some interviewees.

Finally, a follow-up study could be conducted with a different humanoid robot, for example the Wakamaru or ASIMO robot. This could help find out whether the results are transferable to other humanoid robots.

CHAPTER A

Appendix

The Appendix A includes (in order):

1. Interview Consent Form (English)
2. Interview Consent Form (German)
3. Interview Guidelines (English)
4. Interview Guidelines (German)
5. Validation Survey Questions (English)
6. Validation Survey Questions (German)
7. Evaluation Survey Video Story

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Information About the Interview

Thank you for your participation as an interviewee in the master thesis of Sarah Fischer. The interview will take about 30 minutes and includes questions about videos, in which the robot Pepper can be seen having a conversation with a human. The videos will be played and afterwards I will ask you several questions about the video, which will be audio recorded. The goal is to gain insight into the reactions of humans to the robot Pepper.

During the interview there are no wrong answers, your honest opinion is what is important. The interview can be paused or stopped at any time. If you decide to stop and not continue with the interview, all of your data collected up to that point will be deleted.

The following personal data will be processed in connection with my academic writing:

- Personal information such as your age, gender, and professional occupation.
- Audio recording of the interview.

The data will not be used to identify individuals but rather to give a general impression of all of the participants in this study. Personal and contact data are separated from interview and analysis data and stored separately. Pseudonymous references will be created to store transcriptions of the interview and analysis in anonymized form. These anonymous references will be used to remove the data in case you decide to withdraw from the project. Anonymized data and results will be included in the master thesis of Sarah Fischer and might also be submitted for publication or presentation at scientific venues in Austria or internationally. The data will be stored on safe media for a maximum of 5 years.

If you have any questions or want to withdraw from the interview at a later date, please contact Sarah Fischer (sarah.fischer@tuwien.ac.at).

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Participation Consent Form

By signing the document, you agree to all the following points:

- You have read and understood the information given above.
- You understand that the interview will be audio recorded.
- You are voluntarily participating and can withdraw at any time.
- You understand that, should you wish to no longer participate it will have no negative consequences.
- You can refuse to answer any question without justification or negative consequences.

_____ _____

Date Signature of the Interviewee

_____

Signature of the Interviewer

Data Consent Form

By signing the document, you agree to all the following points:

- You allow the researcher to store and process data such as the audio recording of the interview.
- You agree for your anonymized data to be used for the purpose of the master thesis and publications of Sarah Fischer.
- You have been informed that you may withdraw your data consent at any time with effect for the future with no negative consequences.
- You may send your declaration of withdrawal to Sarah Fischer (sarah.fischer@tuwien.ac.at).

_____ _____

Date Signature of the Interviewee

_____

Signature of the Interviewer

A signed and dated copy of this consent form is for you to keep.

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Informationen über das Interview

Danke, dass Sie an diesem Interview im Zuge der Masterarbeit von Sarah Fischer teilnehmen. Das Interview wird ungefähr 30 Minuten dauern und beinhaltet Fragen über Videos, in denen der Roboter Pepper sich mit einem Menschen unterhält. Die Videos werden vorgespielt und danach werde ich Ihnen einige Fragen zu dem Video stellen, welche mittels Audioaufnahme aufgezeichnet werden. Ziel ist es Einblick in die Reaktionen von Menschen auf den Roboter Pepper zu erlangen.

Während des Interviews gibt es keine falschen Antworten, Ihre ehrliche Meinung ist worauf es ankommt. Das Interview kann zu jedem Zeitpunkt pausiert oder gestoppt werden. Falls Sie sich dazu entschließen das Interview zu stoppen und nicht weiterzuführen, dann werden alle Ihre Daten, die bis zu diesem Zeitpunkt gesammelt wurden, gelöscht.

Die folgenden persönlichen Daten werden in Zusammenhang mit meinen akademischen Arbeiten bearbeitet:

- Persönliche Informationen wie Ihr Alter, Geschlecht und Beruf.
- Audioaufnahme des Interviews.

Die Daten werden nicht dazu verwendet Individuen zu identifizieren, sondern um einen generellen Eindruck über die Menschen, die an dieser Studie teilnehmen zu vermitteln. Persönliche Daten und Kontaktdaten werden von der Aufnahme des Interviews und den Analysedaten getrennt und separat aufbewahrt.
Um die Transkripte und die Analyse des Interviews lagern zu können, werden Pseudonyme erstellt. Diese anonymen Referenzen werden auch verwendet, um Ihre Daten zu löschen, falls Sie im Nachhinein von dem Interview zurücktreten möchten. Anonymisierte Daten und Ergebnisse werden in der Masterarbeit von Sarah Fischer vorkommen und werden möglicherweise auch zur Publikation oder Präsentation bei wissenschaftlichen Verlagen oder Vorträgen (sowohl in Österreich als auch international) eingereicht. Die Daten werden auf sicheren Medien für maximal 5 Jahre aufbewahrt.

Falls Sie irgendwelche Fragen haben oder zu einem späteren Zeitpunkt von dem Interview zurücktreten wollen, kontaktieren Sie bitte Sarah Fischer (sarah.fischer@tuwien.ac.at).

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Teilnahme Einwilligungserklärung

Mit Ihrer Unterschrift auf diesem Dokument stimmen Sie den folgenden Punkten zu:

- Sie haben die obigen Informationen gelesen und verstanden.
- Sie haben verstanden, dass das Interview mit einem Audiogerät aufgezeichnet wird.
- Sie nehmen freiwillig an dem Interview teil und können jederzeit davon zurücktreten.
- Sie verstehen, dass, falls Sie nicht mehr teilnehmen möchten, keine negativen Konsequenzen daraus für Sie entstehen.
- Sie können sich jederzeit weigern eine Frage zu beantworten, ohne Angabe von Gründen und ohne dass Ihnen dadurch negative Konsequenzen entstehen.

_____ _____

Datum Unterschrift der befragten Person

_____

Unterschrift des Interviewers

Einwilligungserklärung zur Datenverarbeitung

Mit Ihrer Unterschrift auf diesem Dokument stimmen Sie den folgenden Punkten zu:

- Sie erlauben der Wissenschaftlerin Ihre Daten, wie z.B. die Audioaufnahmen des Interviews, zu lagern und zu verarbeiten.
- Sie stimmen zu, dass Ihre anonymisierten Daten in der Masterarbeit sowie Publikationen von Sarah Fischer verwendet werden.
- Sie wurden darüber informiert, dass Sie Ihre Einwilligungserklärung jederzeit zurücknehmen können, ohne dass Ihnen dadurch negative Konsequenzen entstehen.
- Sie können ihre Rücktrittserklärung an Sarah Fischer (sarah.fischer@tuwien.ac.at) senden.

_____ _____

Datum Unterschrift der befragten Person

_____

Unterschrift des Interviewers

Eine unterschriebene und datierte Ausgabe dieser Einwilligungserklärung verbleibt bei Ihnen.

Interview Guidelines

Description

- Please describe in your own words what happened in the video.
    - Describe the robot in your own words.
        - What did you notice about the robot?
        - What impression does the robot make on you?
        - Is the robot likable, or not, why?
        - Would you like to interact with the robot? If so, how do you picture that to be like?

Feelings

- How did you feel during the video?
    - Did you feel uncomfortable at any point, why?
- What did you like / not like and why?
    - Spoken words, Movements, Mood, Attitude, etc.?

- Is there anything else, which was not talked about already, that you want to mentioned / talk about?

Interview Leitfaden

Beschreibung

- Bitte beschreiben Sie in Ihren eigenen Worten was in dem Video passiert.
    - Beschreiben sie den Roboter in Ihren eigenen Worten.
        - Was ist Ihnen an dem Roboter aufgefallen?
        - Was für einen Eindruck macht der Roboter auf Sie?
        - Ist der Roboter sympathisch, unsympathisch, wieso?
        - Würden Sie gerne mal mit dem Roboter interagieren? Wenn ja, wie stellen Sie sich das vor?

Gefühle

- Wie haben sie sich während des Videos gefühlt?
    - Haben Sie sich zu einem Zeitpunkt unwohl gefühlt, wieso?
- Was hat Ihnen gefallen / nicht gefallen und wieso?
    - Gesagtes, Bewegungen, Stimmung, etc.?

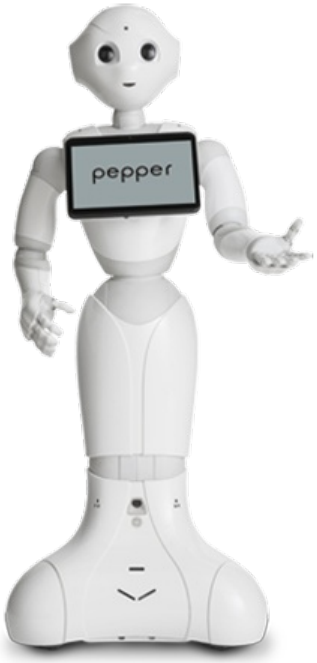
- Möchten Sie noch irgendetwas ansprechen / hinzufügen, dass nicht besprochen wurde?

Online Survey

TU WIEN Informatics

The language can be changed in the top right. Die Sprache kann rechts oben geändert werden.

Thank you for taking part in this online survey! This study is a part of Sarah Fischer's master thesis and wants to investigate how humans perceive and understand the non-verbal behavior of the robot Pepper. The survey consists of general questions and 5 videos, which only take about 40 seconds each. Each video will be followed by 23 descriptive words on a scale, on which the robot in the video is supposed to be rated. The completion of the survey will take about 10 - 15 minutes. Some of the videos are similar on purpose. There are no wrong answers, your personal opinion is what matters.



Please make sure your sound is turned on, so you can hear the voices.

If you have any questions about the study or your rights as a participant, please contact Sarah Fischer (sarah.fischer@tuwien.ac.at).

There are 14 questions in this survey.

This survey is anonymous.

The record of your survey responses does not contain any identifying information about you, unless a specific survey question explicitly asked for it.

If you used an identifying access code to access this survey, please rest assured that this code will not be stored together with your responses. It is managed in a separate database and will only be updated to indicate whether you did (or did not) complete this survey. There is no way of matching identification access codes with survey responses.

General information

*Country you have spent most of your life in

Choose one of the following answers

Austria

*
- I participate in this study voluntarily.
- I understand that I can stop the study at any point, without giving a reason and this will lead to the answers, which I have given to this point, not being a part of the study.
- I understand that this study only collects anonymized and non-identifiable data.
- I understand that my answers, given in this study, will be saved and used in the master thesis of Sarah Fischer.

If you proceed with the survey, you agree that you have read, understood and agree with the statements above.

I confirm that I am 18 years old or older.

✔ Yes	⊘ No

*Age

Choose one of the following answers

Please choose...

*Gender

Choose one of the following answers

- Male
- Female
- Non-Binary
- Other
- Do not want to disclose

*Highest level of education

Choose one of the following answers

- Minimum compulsory schooling
- Apprenticeship
- High School
- College / University
- Other:

In which field?

Are you currently a student?

✔ Yes	⊘ No

*Please rate your experience with robots:

I have never seen a movie / film with a robot	
	○ 1
	○ 2
	○ 3
	○ 4
	○ 5
I work with robots	

Video and questions

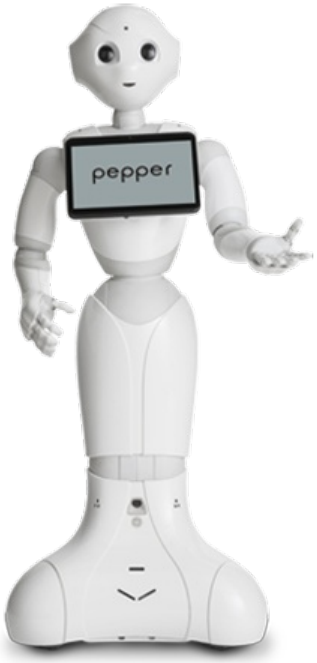
Please make sure your sound is on so you can hear the voices!

Onlineumfrage

TU WIEN Informatics

Die Sprache kann rechts oben geändert werden. The language can be changed in the top right.

Danke, dass Sie an dieser Onlineumfrage teilnehmen! Diese Studie ist Teil der Masterarbeit von Sarah Fischer und möchte erforschen, wie Menschen das nonverbale Verhalten des Roboters Pepper wahrnehmen und verstehen. Die Umfrage besteht aus allgemeinen Fragen und 5 Videos, die jeweils ungefähr 40 Sekunden dauern. Zu jedem Video gibt es anschließend 23 beschreibende Worte auf einer Skala, bezüglich welcher der Roboter in dem Video zu bewerten ist. Die Beantwortung der Umfrage dauert ca. 10 - 15 Minuten. Manche der Videos sind absichtlich ähnlich. Es gibt keine falschen Antworten, es geht um Ihre persönliche Meinung.



Bitte stellen Sie sicher, dass Ihr Ton eingeschaltet ist, damit Sie die Stimmen hören können.

Sie können bei Fragen zu dieser Studie sowie zu Ihren Rechten als Teilnehmer / Teilnehmerin jederzeit Sarah Fischer (sarah.fischer@tuwien.ac.at) kontaktieren.

In dieser Umfrage sind 14 Fragen enthalten.

Dies ist eine anonyme Umfrage.

In den Umfrageantworten werden keine persönlichen Informationen über Sie gespeichert, es sei denn, in einer Frage wird explizit danach gefragt.

Wenn Sie für diese Umfrage einen Zugangscode benutzt haben, so können Sie sicher sein, dass der Zugangsschlüssel nicht zusammen mit den Daten abgespeichert wurde. Er wird in einer getrennten Tabelle aufbewahrt und nur aktualisiert, um zu speichern, ob Sie diese Umfrage abgeschlossen haben oder nicht. Es gibt keinen Weg, die Zugangscodes mit den Umfrageergebnissen zusammenzuführen.

Allgemeine Informationen

*
- Ich nehme freiwillig an dieser Studie teil.
- Ich habe verstanden, dass ich die Studie zu jedem Zeitpunkt, ohne Angabe von Gründen, abbrechen kann, was dazu führt, dass meine bis dahin gegeben Antworten nicht in die Studie aufgenommen werden.
- Ich habe verstanden, dass in dieser Studie nur anonymisierte und nicht identifizierbare Informationen gesammelt werden.
- Ich habe verstanden, dass die Antworten, die ich gebe, gespeichert und in der Masterarbeit von Sarah Fischer verwendet werden.

Mit meiner Teilnahme an der Befragung bestätige ich, dass ich alle soeben genannten Punkte gelesen und verstanden habe und in allen Punkten meine Zustimmung gebe.

Ich bestätige, dass ich 18 Jahre alt, oder älter bin:

✔ Ja	⊘ Nein

*Land in dem Sie den größten Teil Ihres Lebens verbracht haben

Bitte wählen Sie eine der folgenden Antworten:

Austria

*Alter

Bitte wählen Sie eine der folgenden Antworten:

Bitte auswählen..

*Geschlecht

Bitte wählen Sie eine der folgenden Antworten:

- Männlich
- Weiblich
- Non-Binary
- Andere
- Möchte ich nicht angeben

*Höchster abgeschlossener Bildungsgrad

Bitte wählen Sie eine der folgenden Antworten:

- Pflichtschule
- Lehre
- Matura
- Hochschule / Universität
- Sonstiges:

In welchem Fachbereich?

Sind Sie derzeit Student oder Studentin?

✔ Ja	⊘ Nein

*Bitte geben Sie an wie viel Erfahrung Sie mit Robotern haben:

	1	2	3	4	5
Ich habe noch nie Filme / Videos mit Robotern gesehen	○	○	○	○	○
Ich arbeite mit Robotern	○	○	○	○	○

Video und Fragen

Bitte stellen Sie sicher, dass ihr Ton an ist damit Sie die Stimmen hören können!

✱The robot wants to indicate to me:

Check all that apply

- [ ] Attention / Interest
- [ ] Disagreement
- [ ] Defensiveness / Disinterest
- [ ] Agreement
- [ ] Other:

✱Please describe the robot's movements in a few words.

The robot

For example: The robot is waving

✱Der Roboter will mir Folgendes zu verstehen geben:

❗ Bitte wählen Sie die zutreffenden Antworten aus:

- [ ] Widerspruch
- [ ] Defensive / Desinteresse
- [ ] Aufmerksamkeit / Interesse
- [ ] Zustimmung
- [ ] Sonstiges:

✱Bitte beschreiben Sie die Bewegung des Roboters in wenigen Worten.

Der Roboter

❓ Beispiel:

Der Roboter winkt

Video und Fragen

Falls das Video steckengeblieben ist, klicken Sie bitte noch einmal auf den Play Button und schauen sich das Video einmal vollständig an.

Evaluation Survey- Video Stories

Robot Talking

Pepper: Hey Sarah, do you wanna hear a story?

Sarah: Sure.

Pepper: A Dog is walking home with a piece of meat in his mouth. On his way home he crosses a river and looks into the water. He mistakes his own reflection for another Dog and wants his meat also. But as he opens his mouth, the meat falls into the river and is never seen again. End of story. Do you want to hear another one?

Sarah: No thanks.

Robot Listening

Sarah: Hey Pepper, do you wanna hear a story?

Pepper: Sure

Sarah: Ok so, a thirsty crow comes across a pitcher full of water. But when she tries to drink from it the water is too low and she can't reach it. She keeps trying, but eventually she just gives up and thinks she will go thirsty. But suddenly she comes up with an idea! She keeps dropping pebbles into the pitcher until the water rises and she is finally able to drink. End of story. Do you want to hear another story?

Pepper: No thanks.

Adapted from this English translation of Aesop's fables:

https://www.imagineforest.com/blog/life-lessons-aesops-fables/ (last accessed: 04.06.2021)

CHAPTER B

Appendix

The Appendix B includes (in order):

1. The results from Mauchly's test and the repeated measures ANOVA done over the individual questions of each of the Godspeed items of the robot-talking condition in Table B.1
2. The results from the paired t-test done over the individual questions of each of the Godspeed items of the robot-listening condition. Table B.2
3. Bar charts describing the results from some of the individual questions of the robot-talking condition in Figure B.1 and Figure B.2
4. Bar charts describing the results from some of the individual questions of the robot-listening condition in Figure B.3 and Figure B.4

Semantic Scale by Godspeed Item	Mauchly's $X^2(2)$	Mauchly's p	ANOVA F(2, 106)	ANOVA p	α'
Anthropomorphism					
Fake/Natural	0.531	0.767	4.357	0.015	0.01
Machinelike/Humanlike	0.065	0.968	4.25	0.17	0.01
Unconscious/Conscious	0.649	0.723	3.705	0.028	0.01
Artificial/Lifelike	1.573	0.455	4.74	0.011	0.01
Moving rigidly/Moving elegantly	0.543	0.762	1.156	0.319	0.01
Animacy					
Dead/Alive	5.666	0.056	7.179	0.001**	0.008
Stagnant/Lively	1.391	0.499	7.178	0.001**	0.008
Mechanical/Organic	1.649	0.439	3.183	0.045	0.008
Artificial/Lifelike	1.573	0.455	4.74	0.011	0.008
Inert/Interactive	2.205	0.332	2.886	0.6	0.008
Apathetic/Responsive	3.761	0.153	6.638	0.002*	0.008
Likeability					
Dislike/Like	3.471	0.176	1.387	0.254	0.01
Unfriendly/Friendly	0.28	0.869	2.461	0.09	0.01
Unkind/Kind	0.356	0.837	3.067	0.051	0.01
Unpleasant/Pleasant	0.735	0.693	0.886	0.415	0.01
Awful/Nice	1.198	0.549	2.23	0.113	0.01
Perceived Intelligence					
Incompetent/Competent	8.315	0.016	4.454	0.014	0.01
Ignorant/Knowledgeable	0.135	0.935	0.428	0.653	0.01
Irresponsible/Responsible	9.291	0.01	0.445	0.642	0.01
Unintelligent/Intelligent	0.11	0.946	1.72	0.184	0.01
Foolish/Sensible	0.165	0.921	1.166	0.316	0.01
Perceived Safety					
Anxious/Relaxed	7.656	0.022	3.118	0.048	0.017
Calm/Agitated	4.925	0.085	3.267	0.042	0.017
Quiescent/Surprised	3.106	0.212	2.866	0.061	0.017

Table B.1: Results from Mauchly's test and the repeated measures ANOVA done over the individual questions of each of the Godspeed items of the robot-talking condition. The last column displays the α value that was newly calculated with the Bonferroni correction. Statistical significance is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

Semantic Scale by Godspeed Item	Non-Exag.(M)	Non-Exag.(SD)	Exag.(M)	Exag.(SD)	t(53)	p	α'
Anthropomorphism							
Fake/Natural	2.26	0.894	2.33	1.116	-0.522	0.604	0.01
Machinelike/Humanlike	1.96	0.971	1.98	0.961	-0.139	0.890	0.01
Unconscious/Conscious	2.43	0.944	2.70	1.002	-2.326	0.024	0.01
Artificial/Lifelike	2.04	0.910	2.30	1.127	-1.847	0.070	0.01
Moving rigidly/Moving elegantly	1.94	0.940	2.19	1.083	-1.563	0.124	0.01
Animacy							
Dead/Alive	2.85	0.960	3.02	1.141	-1.456	0.151	0.0083
Stagnant/Lively	2.89	1.093	3.31	1.043	-2.961	0.005*	0.0083
Mechanical/Organic	1.94	0.940	2.02	0.942	-0.562	0.576	0.0083
Artificial/Lifelike	2.04	0.910	2.30	1.127	-1.847	0.070	0.0083
Inert/Interactive	3.07	1.025	3.31	1.195	-1.788	0.079	0.0083
Apathetic/Responsive	3.28	0.856	3.72	0.940	-3.385	0.001**	0.0083
Likeability							
Dislike/Like	3.11	1.058	3.48	1.077	-2.428	0.019	0.01
Unfriendly/Friendly	3.63	0.896	3.93	0.908	-2.172	0.034	0.01
Unkind/Kind	3.69	0.773	3.87	0.754	-1.604	0.115	0.01
Unpleasant/Pleasant	3.15	0.940	3.13	0.953	0.125	0.901	0.01
Awful/Nice	3.61	0.878	3.70	1.002	-0.742	0.461	0.01
Perceived Intelligence							
Incompetent/Competent	3.19	0.803	3.13	0.933	0.504	0.617	0.01
Ignorant/Knowledgeable	3.00	0.700	3.09	0.807	-1.043	0.301	0.01
Irresponsible/Responsible	3.11	0.538	3.24	0.581	-1.547	0.128	0.01
Unintelligent/Intelligent	3.06	0.960	3.20	1.035	-1.241	0.220	0.01
Foolish/Sensible	3.31	0.577	3.33	0.727	-0.159	0.875	0.01
Perceived Safety							
Anxious/Relaxed	3.43	1.021	3.48	1.005	-0.454	0.652	0.0167
Calm/Agitated	2.37	0.996	3.15	1.219	-5.071	0.000**	0.0167
Quiescent/Surprised	2.89	1.058	3.39	0.998	-2.860	0.006*	0.0167

Table B.2: Results from paired t-test done over the individual questions of each of the Godspeed items of the robot-listening condition. The last column displays the α value that was newly calculated with the Bonferroni correction. Statistical significance is marked with * ($p \leq 0.01$) and ** ($p \leq 0.001$).

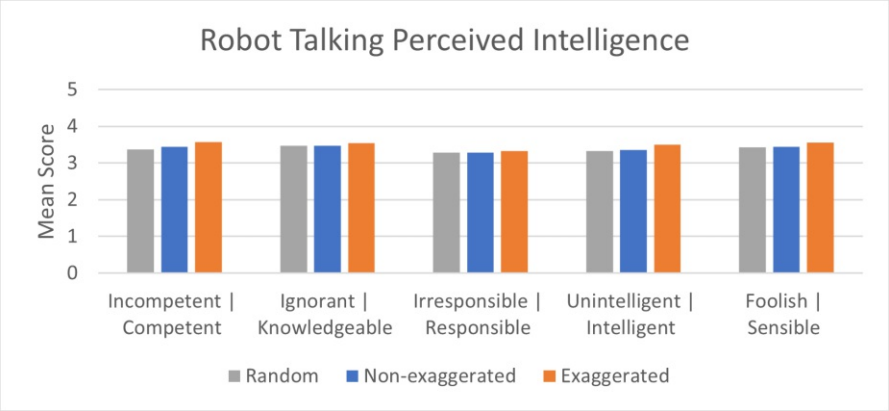


Figure B.1: Charts describing the results for the individual questions of the Godspeed item perceived intelligence, for the robot-talking condition. The y-axes represent the mean values of the Godspeed score and the x-axes show the different questions (semantic differential scales).

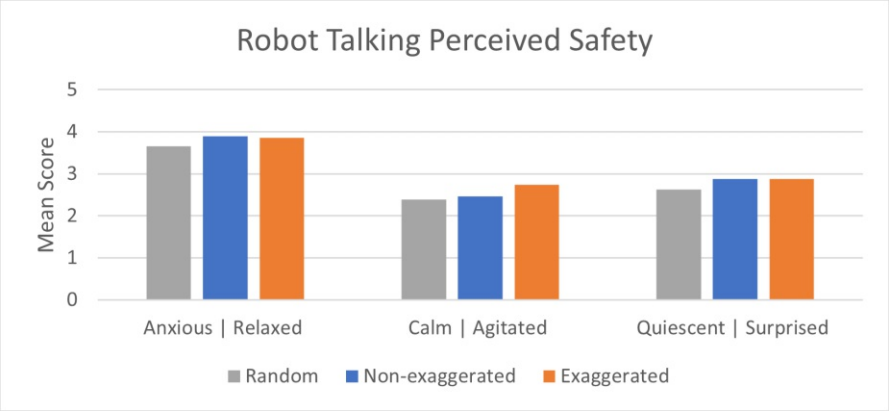


Figure B.2: Charts describing the results for the individual questions of the Godspeed item perceived safety, for the robot-talking condition. The y-axes represent the mean values of the Godspeed score and the x-axes show the different questions (semantic differential scales).

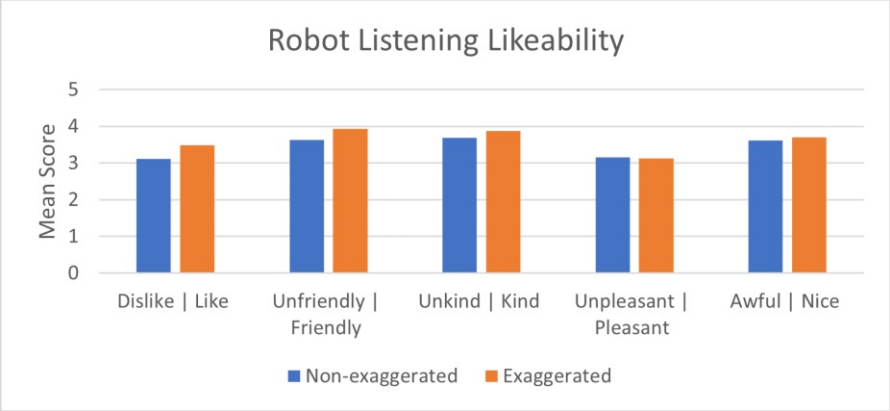


Figure B.3: Charts describing the results for the individual questions of the Godspeed item likeability, for the robot-listening condition. The y-axes represent the mean values of the Godspeed score and the x-axes show the different questions (semantic differential scales).

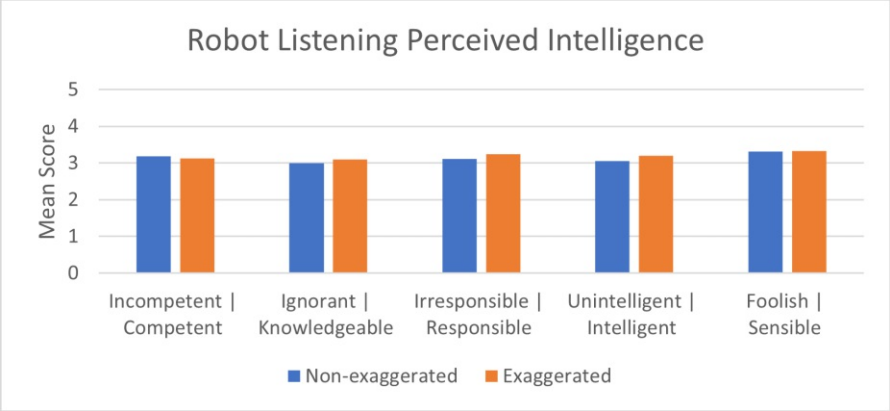


Figure B.4: Charts describing the results for the individual questions of the Godspeed item perceived intelligence, for the robot-listening condition. The y-axes represent the mean values of the Godspeed score and the x-axes show the different questions (semantic differential scales).

List of Figures

List of Tables

Bibliography

[ADGY04] Anthony P. Atkinson, Winand H. Dittrich, Andrew J. Gemmell, and Andrew W. Young. Motion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33(6):717–746, 2004.

[ASF+20] Antonio Andriella, Henrique Siqueira, Di Fu, Sven Magg, Pablo Barros, Stefan Wermter, Carme Torras, and Guillem Alenyà. Do I have a personality? Endowing care robots with context-dependent personality traits. *International Journal of Social Robotics*, 12(5):1–22, 2020.

[Asi42] Isaac Asimov. Runaround. *Astounding Science Fiction*, 1942.

[Aud] AudacityTeam. Audacity. `https://www.audacityteam.org/`. Accessed: 07-06-2021.

[BA95] J. Martin Bland and Douglas G. Altman. Multiple significance tests: the Bonferroni method. *BMJ*, 310(6973):170, 1995.

[BBE+20] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. *Human-robot interaction – An introduction*. Cambridge University Press, 2020.

[BBH20] Adna Bliek, Suna Bensch, and Thomas Hellstrom. How can a robot trigger human backchanneling? In *International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 96–103, 2020.

[BC12] Virginia Braun and Victoria Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*, pages 57–71. American Psychological Association, 2012.

[Bir83] Ray L. Birdwhistell. Background to kinesics. *ETC: A Review of General Semantics*, 40(3):352–361, 1983.

[BKCZ09] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.

[BKE10] Kirsten Bergmann, Stefan Kopp, and Friederike Eyssel. Individualized gesturing outperforms average gesturing – Evaluating gesture production in virtual humans. In *International Conference on Intelligent Virtual Agents (IVA)*, pages 104–117, 2010.

[BKM+21] Franziska Babel, Johannes Kraus, Linda Miller, Matthias Kraus, Nicolas Wagner, Wolfgang Minker, and Martin Baumann. Small talk with a tobot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. *International Journal of Social Robotics*, 13(6):1485–1498, 2021.

[CCM12] Vijay Chidambaram, Yueh Hsuan Chiang, and Bilge Mutlu. Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. In *International Conference on Human-Robot Interaction (HRI)*, pages 293–300, 2012.

[CDFV18] Bart Craenen, Amol Deshmukh, Mary Ellen Foster, and Alessandro Vinciarelli. Do we really like robots that match our personality? the case of Big-five traits, Godspeed scores and robotic gestures. In *International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 626–631, 2018.

[Dar15] Charles Darwin. *The expression of the emotions in man and animals*. University of Chicago Press, 2015.

[Dau03] Kerstin Dautenhahn. Roles and functions of robots in human society: Implications from research in autism therapy. *Robotica*, 21(4):443–452, 2003.

[Dau07] Kerstin Dautenhahn. Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704, 2007.

[DW04] Kerstin Dautenhahn and Iain Werry. Towards interactive robots in autism therapy. *Pragmatics & Cognition*, 12(1):1–35, 2004.

[FEN+04] Shinya Fujie, Yasushi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi. A conversation robot using head gesture recognition as para-linguistic information. In *International Workshop on Robot and Human Interactive Communication*, pages 159–164, 2004.

[FIPC15] Luis A. Fuente, Hannah Ierardi, Michael Pilling, and Nigel T. Crook. Influence of upper body pose mirroring in human-robot interaction. In *International Conference in Social Robotics*, pages 214–223, 2015.

[FND03] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3):143–166, 2003.

[Fre] FreeBSD. Freebsd ports. `https://www.freebsd.org/ports`. Accessed: 07-06-2021.

[FWL+13] Stephen M. Fiore, Travis J. Wiltshire, Emilio J.C. Lobato, Florian G. Jentsch, Wesley H. Huang, and Benjamin Axelrod. Toward understanding social cues and signals in human-robot interaction: Effects of robot gaze and proxemic behavior. *Frontiers in Psychology*, 4(11):1–15, 2013.

[FXh] FXhome. Hitfilm express. `https://fxhome.com/product/hitfilm-express`. Accessed: 07-06-2021.

[Gmb] LimeSurvey GmbH. Limesurvey. `https://www.limesurvey.org/de/`. Accessed: 07-06-2021.

[GS07] Michael A. Goodrich and Alan C. Schultz. Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275, 2007.

[GT12] Michael J. Gielniak and Andrea L. Thomaz. Enhancing interaction through exaggerated motion synthesis. In *International Conference on Human-Robot Interaction (HRI)*, pages 375–382, 2012.

[GZ12] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486–489, 2012.

[HGFT07] N. G. Hockstein, C. G. Gourin, R. A. Faust, and D. J. Terris. A history of robots: From science fiction to surgical robotics. *Journal of Robotic Surgery*, 1(2):113–118, 2007.

[HHM19] Judith A. Hall, Terrence G. Horgan, and Nora A. Murphy. Nonverbal communication. *Annual Review of Psychology*, 70(1):271–294, 2019.

[Hon] Honda. Asimo robot. `https://asimo.honda.com/`. Accessed: 31-08-2021.

[IBM] IBM. SPSS. `https://www.ibm.com/at-de/products/spss-statistics`. Accessed: 08-10-2021.

[IBM14] IBM. *IBM SPSS Statistics 22 Algorithms*. IBM, 2014.

[ILIH10] Carlos T. Ishi, ChaoRan Liu, Hiroshi Ishiguro, and Norihiro Hagita. Head motion during dialogue speech and nod timing control in humanoid robots. In *International Conference on Human-Robot Interaction (HRI)*, pages 293–300, 2010.

[Ind] Mitsubishi Heavy Industries. Wakamaru robot. `https://robots.ieee.org/robots/wakamaru/`. Accessed: 31-08-2021.

[LAN] Video LAN. VLC Player. `https://www.videolan.org/vlc/index.de.html`. Accessed: 07-06-2021.

[LIIH13] Chaoran Liu, Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Generation of nodding, head tilting and gazing for human-robot speech interaction. *International Journal of Humanoid Robotics*, 10(1):1–19, 2013.

[LJN15] Jamy Li, Wendy Ju, and Cliff Nass. Observer perception of dominance and mirroring behavior in human-robot relationships. In *International Conference on Human-Robot Interaction (HRI)*, pages 133–140, 2015.

[LLP21] Knowledge Sourcing Intelligence LLP. *Social Robots Market - Forecasts from 2021 to 2026*. ResearchAndMarkets.com, 2021.

[LR87] John Lasseter and San Rafael. Principles of traditional animation applied to 3D computer animation. In *Conference on Computer Graphics and Interactive Techniques*, pages 35–44, 1987.

[Ltd] Lund Research Ltd. Repeated measures anova. `https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php`. Accessed: 08-10-2021.

[Mav15] Nikolaos Mavridis. A review of verbal and non-verbal human-robot interactive communication. *Robotics and Autonomous Systems*, 63(1):22–35, 2015.

[MF67] Albert Mehrabian and Susan R Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31(3):248, 1967.

[MMK12] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. The Uncanny Valley [From the Field]. *Robotics & Automation Magazine*, 19(2):98–100, 2012. Translation of Masahiro Mori, Bukimi no tani (the uncanny valley), *Energy*, 7(4):33–35, 1970 (in Japanese).

[Mut11] Bilge Mutlu. Designing embodied cues for dialogue with robots. *AI Magazine*, 32(4):17–30, 2011.

[MW67] Albert Mehrabian and Morton Wiener. Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, 6(1):109, 1967.

[Oan18] Stefan O. Oancea. *Four texture algorithms for recognizing early signs of osteoarthritis. Data from the multicenter osteoarthritis study*. Master's Thesis, Technische Universität Wien, 2018.

[OKI+08] Yusuke Okuno, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. Providing route directions: Design of robot's utterance, gesture, and timing. *International Conference on Human-Robot Interaction (HRI)*:53–60, 2008.

[otUni10] Psychology Writing Center of the University of Washington. Part A : Reporting Results of Common Statistical Tests in APA Format. *Statistics*, 5(33):1–8, 2010.

[Pan16] Nikolaos Pandis. The chi-square test. *American Journal of Orthodontics and Dentofacial Orthopedics*, 150(5):898–899, 2016.

[PAR] Inc. PARO Robots U.S. Paro robot. `http://www.parorobots.com/`. Accessed: 31-08-2021.

[PG18] Amit Kumar Pandey and Rodolphe Gelin. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *Robotics and Automation Magazine*, 25(3):40–48, 2018.

[PLC15] Jeong Woo Park, Hui Sung Lee, and Myung Jin Chung. Generation of realistic robot facial expressions for human robot interaction. *Journal of Intelligent and Robotic Systems: Theory and Applications*, 78(3):443–462, 2015.

[REF+20] Danielle Rifinski, Hadas Erel, Adi Feiner, Guy Hoffman, and Oren Zuckerman. Human-human-robot interaction: Robotic object's responsive gestures improve interpersonal evaluation in human interaction. *Human–Computer Interaction*, 36(4):1–27, 2020.

[Roba] Softbank Robotics. Choregraphe. `http://doc.aldebaran.com/2-4/software/choregraphe/index.html`. Accessed: 14-09-2021.

[Robb] Softbank Robotics. Nao robot. `https://www.softbankrobotics.com/emea/de/nao`. Accessed: 31-08-2021.

[Robc] Softbank Robotics. Pepper. `https://www.softbankrobotics.com/emea/en/pepper`. Accessed: 03-07-2020.

[Robd] Softbank Robotics. Pepper's joints. `http://doc.aldebaran.com/2-5/family/pepper_technical/joints`. Accessed: 03-02-2020.

[Robe] Softbank Robotics. Photo of pepper. `https://www.softbankrobotics.com/emea/de/pepper-and-nao-robots-education`. Accessed: 18-08-2021.

[Robf] Softbank Robotics. Softbank robotics. `https://www.softbankrobotics.com/`. Accessed: 12-09-2021.

[RP12] Tiago Ribeiro and Ana Paiva. The illusion of robotic life: Principles and practices of animation for robots. In *International Conference on Human-Robot Interaction (HRI)*, pages 383–390, 2012.

[SER+13] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3):313–323, 2013.

[SG96] Gerard Saucier and Lewis R. Goldberg. The language of personality: Lexical perspectives on the five-factor model. In *The five-factor model of personality: Theoretical perspectives.* Pages 21–50. Guilford Press, 1996.

[SN19] Shane Saunderson and Goldie Nejat. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, 11(4):575–608, 2019.

[STH18] Trenton Schulz, Jim Torresen, and Jo Herstad. Animation techniques in human-robot interaction user studies: A systematic literature review. *arXiv*, 8(2):1–22, 2018.

[SW65] Samuel S. Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3):591–611, 1965.

[TDJ11] Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: Improving robot readability with animation principles. In *International Conference on Human-Robot Interaction (HRI)*, pages 69–76, 2011.

[Tea] Handbreak Team. Handbreak. `https://handbrake.fr/`. Accessed: 30-08-2021.

[Tec] Georgia Tech. Simon robot. `https://robots.ieee.org/robots/simon`. Accessed: 05-07-2020.

[TJ95] Frank. Thomas and Ollie. Johnston. *The Illusion of Life: Disney Animation*. Disney Editions, 1995.

[TMS20] Yunus Terzioglu, Bilge Mutlu, and Erol Sahin. Designing social cues for collaborative robots: The role of gaze and breathing in human-robot collaboration. In *International Conference on Human-Robot Interaction (HRI)*, pages 343–357, 2020.

[To] Statistics How To. F Statistic / F Value: Simple Definition and Interpretation. `https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/WhatisF`. Accessed: 10-10-2021.

[Uni] Kent State University. SPSS Tutorials: Crosstabs. `https://libguides.library.kent.edu/spss/crosstabs`. Accessed: 10-10-2021.

[Wal98] Harald G Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28(6):879–896, 1998.

[Wit85] Andreas Witzel. Das problemzentrierte interview. In *Qualitative Forschung in der Psychologie : Grundfragen, Verfahrensweisen, Anwendungsfelder*, pages 227–255. Beltz, 1985.

[WSD+08] Michael L. Walters, Dag S. Syrdal, Kerstin Dautenhahn, René te Boekhorst, and Kheng Lee Koay. Avoiding the uncanny valley: Robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, 24(2):159–178, 2008.

[WWKD06] Sarah Woods, Michael Walters, Kheng Lee Koay, and Kerstin Dautenhahn. Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach. In *International Workshop on Advanced Motion Control (AMC)*, pages 750–755, 2006.

[XBHN14] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A. Neerincx. Robot mood is contagious: Effects of robot body language in the imitation game. In *Artificial Intelligence Conference*, pages 181–182, 2014.

[ZDL+17] Cristina Zaga, Roelof A.J. De Vries, Jamy Li, Khiet P. Truong, and Vanessa Evers. A simple nod of the head: The effect of minimal robot movements on children's perception of a low-anthropomorphic robot. In *Conference on Human Factors in Computing Systems*, pages 336–341, 2017.

[ZME+09] Massimiliano Zecca, Yu Mizoguchi, K. Endo, F. Iida, Y. Kawabata, Nobutsuna Endo, Kazuko Itoh, and Atsuo Takanishi. Whole body emotion expressions for KOBIAN humanoid robot - Preliminary experiments with different emotional patterns. In *International Workshop on Robot and Human Interactive Communication*, pages 381–386, 2009.