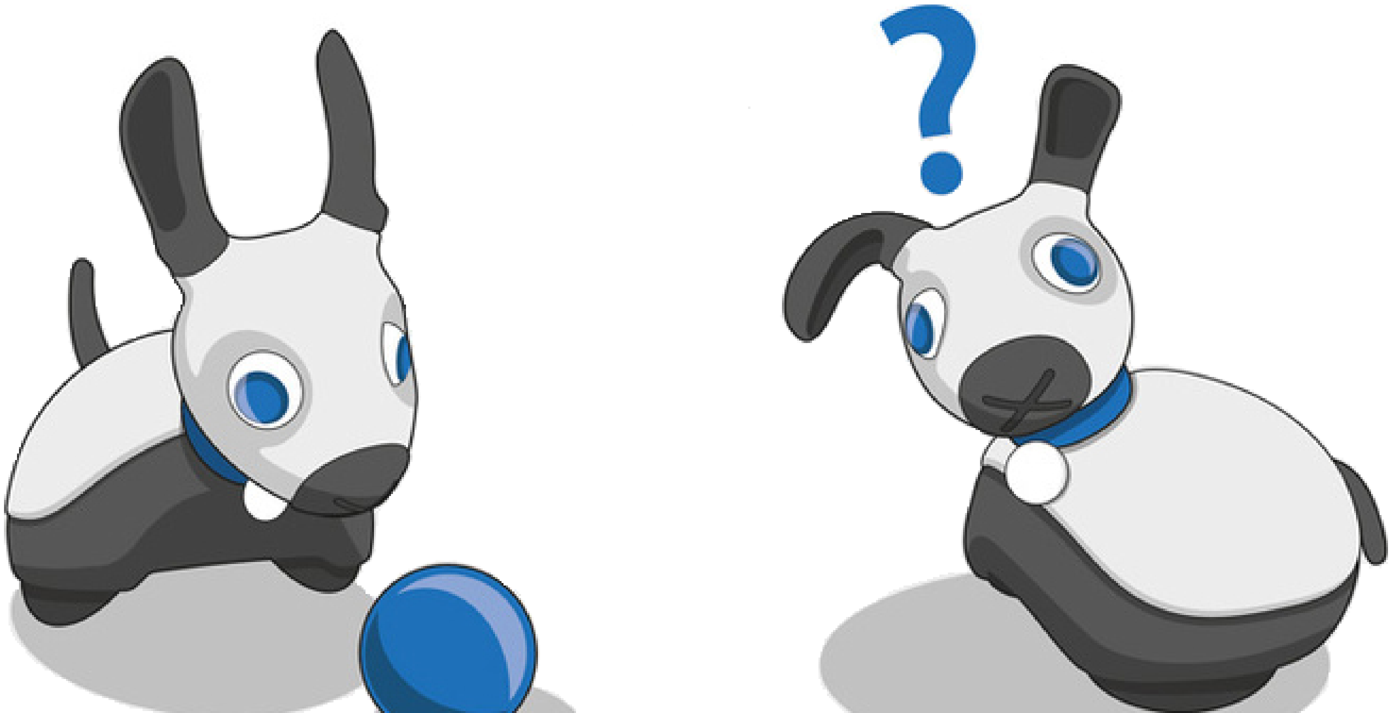
Levels of joint attention in an interaction between humans and an animal-like social robot

M.J.E. van Osch



TU Delft

Levels of joint attention in an interaction between humans and an animal-like social robot

by

M.J.E. van Osch

To obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday December 12, 2023 at 14:30.

Project duration:	September 5, 2022 – December 12, 2023		
Thesis Committee:	Dr. ir. W.P. Brinkman,	TU Delft,	Thesis advisor
	Dr. ir. F. Broz,	TU Delft,	Daily supervisor
	Dr. ir. U.K. Gadiraju,	TU Delft	

TU Delft

Acknowledgements

Although I worked alone on this research project, this thesis would not have come to completion without the support of other people.

I would therefore like to thank my daily supervisor Frank Broz for his feedback and guidance throughout this research. You gave me the things I needed to get started with this research project and were as enthusiastic about my ideas as I was. Our meetings helped shape and guide the direction of this research and I enjoyed bouncing my ideas off you during these meetings. I would also like to thank my thesis supervisor Willem-Paul Brinkman for bringing me in contact with Frank, but also for reminding me of angles I had not considered and correcting my path forward where needed. I also want to thank the lovely people of consequential robotics; you were quick to respond whenever I had a question about the robot or if an issue arose with the robot. Your help with solving a movement issue with the robot is greatly appreciated. Furthermore, I want to thank my friends and family. Thank you for keeping me sane during this research project and helping me relax, especially during the stressful moments.

Lastly, I want to give special thanks to three people. To my best friend Bart, thank you for letting me bounce my ideas off you, for making the time to proofread and give feedback and for allowing me to ramble on whenever things weren't going as well as I hoped. And to my dear parents: thank you for all the support such as providing me with a space to work and perform my experiments, and for pushing me when I needed it most. I know I was not always easy, but I truly think I could not have completed this thesis without your support.

Millen van Osch
Breda, December 2023

Abstract

Joint attention is the shared focus multiple people can have on the same object and it is subconsciously used by humans every day. The simple act of verbally or non-verbally pointing out an object to one another, is a form of joint attention. Its use facilitates human cooperation, such as when someone needs to hand over an object to another person. When robots display those same joint attention behaviours, it could be useful in improving human-robot interactions and perhaps variably so in various levels of jointness. This research investigates the effects of these levels of jointness on the interactions between a human and a robot. This is done by investigating the effect on the person's task performance, the effect on their mental model of the robot, and the effect on their perception of the robot as its own entity with its own mind. To this end, this research defines four levels of joint attention and designs an experiment that makes use of these four levels of joint attention. To perform this experiment, a system capable of establishing joint attention has been developed. The system is divided into two parts; hardware and software tools & applications which have not been developed in this project but are used as is, and software which has been actively developed in this research. In this experiment, participants played three guessing games with a robot. Task performance was measured by the amount of time and hints needed, as well as the accuracy of a participant. The participant's mental model and perception of the robot as its own entity with its own mind were measured using a questionnaire. This research found no significant effect on the mental model towards the robot and the perception of whether the robot is its own entity with its own mind, although it is *not* concluded that such an effect *cannot* be found. This research did find a significant effect on task performance; higher levels of joint attention lead to faster task completion with less hints needed. But interestingly, this did not necessarily lead to a higher accuracy. The system introduced by this research was originally intended to help children with autism spectrum disorder learn joint attention skills, but due to ethical and time constraints, the system was tested with and the experiment was performed with adults instead. Due to ethical constraints, the participants in this research were not asked to disclose whether they had an autism spectrum disorder diagnosis or not. However, future research could replicate this study with the intended target group.

Contents

List of Tables

List of Figures

1 Introduction

Using words is not the only way humans communicate with each other. Humans also use non-verbal communication such as eye-gaze, gestures and body language to convey mental states or augment verbal communications [1, 2]. Eye-gaze is an especially important signal, as evidence has suggested humans have dedicated hard-wired pathways in their brains to interpret these signals, as opposed to other non-verbal cues such as body posture [3]. Eye-gaze is also important when referring to the environment: people look at objects around them before they name them [4]. People can use their social partner's gaze to predict their partner's next verbal object reference and respond more quickly to that reference [5]. One behaviour that makes use of gaze behaviour is joint attention. Joint attention is when you and someone else share attention on the same object. This behaviour is something that humans use throughout their entire lives. A simple example would be a parent reading a children's book to their child. The parent can point out a picture in the book to the child, and the child can confirm they saw it. This phenomena of both parties being aware of the picture in the book and both having their focus on it is an example of joint attention.

Human behaviour like this could be used to improve the interaction between humans and robots. A study by Boucher et al. [5] found that when robots make use of gaze behaviour when interacting with humans, the human's accuracy and response times improved. Adding gaze behaviour to robots is not only a matter of performance, but also preference; according to a study by Willemse and Wykowska [6], humans working with robots prefer it when robots follow their gaze.

Much of the research and development of human robot interaction is done with human adults in mind. But robots can also have a positive effect for children. Robots have been proven to be effective at promoting engagement and reducing stress for children. Furthermore, for children who are not as responsive to conventional human interaction, robots may be able to connect with them [7]. This opens a way to help children with robots where human help may not be as effective. Wainer et al. [8] showed that it is possible to help children with Autism Spectrum Disorder (ASD) who have impaired interaction skills learn and develop these social skills. Robots that make use of human gaze behaviour definitely have potential. A limitation however, is that most robots capable of joint attention and gaze tracking in a social setting are human-like. The problem with this however is that people have expectations of what such a robot should be able to do and modern day technology can not always meet these expectations. When a robot is animal-like or even machine-like instead of human-like, these expectations are different and often lower [9]. Most important of all, since the expectations are lower, they are more likely to be able to be met by current day technology. However, these robots are not always capable of gaze tracking and joint attention, like the MiRo-E for example.

1.1. Research Questions

There are studies that try to define different levels of joint attention like in the work of Siposova et al. [10], but there are not many studies that research the effects of different levels of joint attention. In the study by Yonezawa et al. they investigated whether joint attention had an impact on a human's mental model. They studied if the use of joint attention caused an increase in favourable feeling towards the robot [11]. But while this study did make use of joint attention, it was an all or nothing scenario. They either used it or they didn't, there was no investigation into the levels of joint attention. Joint attention makes use of gaze behaviour, and in a study by Boucher et al. [5] they used gaze cues of differing clarity. They investigated if humans performed better the more clear the gaze cues were. In this research it was investigated what effect the levels of joint attention have on the interaction with a robot. Hence the research question was as follows:

Main RQ: What is the effect of the level of joint attention on the interaction with a human?

In order to better answer the research question it was divided into a few sub-questions that needed to be answered first. There are studies that investigated whether gaze behaviour improved performance like a study by Boucher et al. [5] mentioned above. But there are not many if not no studies that research the effects of different levels of joint attention on task performance. In this research task performance was seen as part of the interaction with the robot and joint attention makes use of gaze behaviour, so the first sub research question was:

SRQ 1: What is the effect of the level of joint attention on task performance?

In this research, a human's perception of the robot is deemed part of the human's mental model of the robot. This perception of the robot also has an impact on the interaction between a human and a robot [12]. Mental model includes subjective measures such as whether a participant likes the robot or trusts the robot to perform its job. This lead to the second sub research question:

SRQ 2: What is the effect of the level of joint attention on the human's mental model of the robot?

Whether a robot is perceived as its own entity with its own mind or not has an influence on how well humans respond to non-verbal communication. This perception is largely affected by if the robot is perceived as a real, intentional being [13]. The perception of the robot as a real and intentional being is in turn influenced by whether or not the robot displays mutual gaze behaviour [14]. So the last sub research question was:

SRQ 3: What is the effect of the level of joint attention on the perception of a robot as its own entity with its own mind?

1.2. Contribution

In this research project an experiment using joint attention has been designed. The experiment made use of different levels of joint attention, for which no single, global definition exists. After reading multiple studies discussing levels of joint attention, this paper defined four levels of joint attention. In order to perform the experiment, software has been developed and a system was designed that can make use of the visual data from an animal-like robot called MiRo-E and from eye tracking glasses to detect the mutual gaze necessary for joint attention. The processing of the robot's visual data is flexible enough so that this system can be adapted to any robot which has built-in cameras and makes use of ROS. The system was used to perform an experiment where the participant played a game with the robot. In order to create the specific movements of the robot for the game, a new tool has also been developed. This allows researchers to easily adjust existing movements or create new ones. The system also makes use of a neural network to detect the robot. While the neural network itself makes use of TensorFlow's implementation of VGG16, the software surrounding the neural network to use the data from the neural network to detect mutual gaze has been developed in this project. Furthermore, some software to guide and help facilitate the training of the network has also been developed in this project. This is done to facilitate reusing this system to experiment with different robots and environments.

1.3. Approach

First existing studies and literature were studied in order to shape the experiment. The background and related work is discussed in more detail in chapter 2. Next, an experiment needed to be designed and conducted. The experimental design is all explained in chapter 3. In order to perform this experiment a system needed to be designed. The system and its design choices are expanded upon in-depth in chapter 4. To properly answer the sub-questions and the main research question, the results were required to be analysed and evaluated. Further details about the experiment's results and the evaluation can be found in chapter 5. After the analysis and evaluation, the results of the analysis and limitations of the work done in this thesis are discussed in chapter 6. Finally this thesis ends with a the conclusions and a future work section containing recommendations and suggestions to build further upon this research.

2 Background & Related work

Research into gaze behaviour in human-robot interaction (HRI) is not something new, roboticists began implementing gaze behaviour into their systems more than 20 years ago [14]. Using robots in this kind of research has its advantages: when used in naturalistic experimental setups it can be ecologically valid and offer excellent control over the experiment [15]. In this paper the focus lies on using joint attention in human-robot interaction.

2.1. Joint attention

2.1.1. Mutual and Referential gaze

Joint attention means sharing attention or focus on an object with someone else, and mutual and referential gaze are powerful tools to establish joint attention. So before joint attention is discussed, mutual gaze and referential gaze need to be covered first.

Mutual gaze

Mutual gaze is when you use your gaze to look someone in the eyes, and they use their gaze to look into your eyes. Mutual gaze is important in the perception of others emotional state and a key part of social communication, furthermore it is a foundational skill which is necessary in the development of joint attention [16]. Mutual gaze has even been described as ”the most powerful mode of establishing a communicative link between humans” [17].

Referential gaze

Referential gaze is when one uses their gaze to refer to an object and another interprets this gaze as referential [18]. Referential gaze is critical in social learning, communication and the inference of mental states [19]. Referential gaze is also a non-verbal disambiguation tool. When humans refer to another object but there is uncertainty about which object is referred to, a human’s gaze is an important and flexible cue to disambiguate this uncertainty [20, 21].

2.1.2. Levels of joint attention

While joint attention has been described as merely sharing attention on an object, it is a little less simple than that. Joint attention has multiple levels and many studies refer to them with different names and terms. Most studies mark the different levels by how aware participants of joint attention are of the others’ attention. Most studies define the lowest level as both participants merely having their attention on the same object, but are not aware the other has their attention on the same object [10, 22]. The highest level of joint attention is marked by both participants being aware the other is aware the attention is shared [10, 14]. After reading through the works of Candland [22], Siposova and Carpenter [10] and Admoni and Scassellati [14], four levels of joint attention were newly defined for this paper. The first level was the starting level of the experiment, the other three levels were actively used in the experiment; these were the levels of the independent variable of the experiment. Figures 2.1 through 2.4 contain visualisations of these levels.

1. **Common attention**

   This is the lowest level of joint attention. In this level of joint attention two people have their attention on the same object, but are not aware of the other. For example when two neighbours watch the same car on the street, but are not aware of each other. This level is not a level of the independent variable of the experiment, but is instead the starting level. This is because there is no active interaction between the participants of the joint attention. By nature of the experiment however, this level was already achieved. See chapter 3 for more details about the experiment.

2. **Monitoring attention**

   The next level of joint attention. This level of joint attention is largely the same as the previous level except that one party is aware that the other sees the same thing. An example of this would be gaze following. This is essentially a form of monitoring attention. This level was used as the lowest level of the independent variable.

3. **Mutual attention**

   In mutual attention both parties are aware that both parties have their attention on the same object, but none or only one party is aware that the other party is aware that both parties have their attention on the same object. An example of this would be a concert. Person A looks at the band on stage, person B next to A is also looking at the band on stage. A is aware that B is looking at the band, and B is aware that A is looking at the band. But while A may be aware that B is aware that A is looking at the band, B may not be aware that A is aware that B is looking at the band. The only thing that is certain is that both parties are looking at the band and know the other is looking at the band. This may seem like a minor detail, but it is actually the defining feature of the highest level of joint attention. This level was used as the middle level of the independent variable.

4. **Shared attention**

   This is the highest level of joint attention. In this level both parties are looking at the same object, both parties are aware the other looks at the same object, but also that both parties are aware the other party is aware that they are looking at the same object. The difference between this level and the previous one is confirmation, which requires communication. This level was used as the highest level of the independent variable.

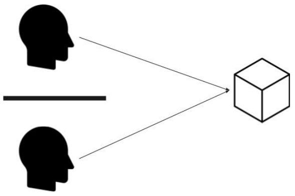


Figure 2.1: Common attention

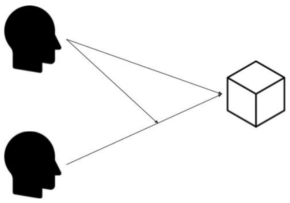


Figure 2.2: Monitoring attention

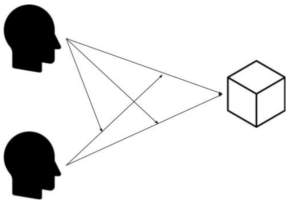


Figure 2.3: Mutual attention

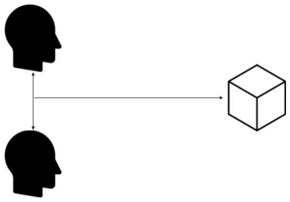


Figure 2.4: Shared attention

2.1.3. Phases

The establishment of joint attention consists of multiple phases. First is getting the attention of the other party with whom one tries to establish joint attention. This can be done with verbal cues such as calling someone by their name or non-verbal cues such as mutual gaze. Following this is focusing one's attention on the object on which they want to establish joint attention. This can also be done with verbal cues by calling it out, i.e. "look at that red apple". Or by non verbal cues ranging from simply pointing to using referential gaze. Humans are quite good at following eye lines, and start exhibiting this trait in their first year [23]. And last is confirming the joint attention is established. This can also be done verbally, or non-verbal like with establishing mutual gaze again. In this research only non-verbal cues were used.

The phases and levels are linked, when two people are in the first level as defined above and one of the two uses mutual gaze, this then raises the joint attention from common attention to monitoring attention. If referential gaze is used after the mutual gaze is established, the level rises again to mutual attention. If the last step of mutual gaze is performed again, shared attention is achieved. See Appendix A for a full breakdown.

2.2. Joint attention between humans and living beings

2.2.1. Joint attention in Human-Human interaction

Humans use joint attention throughout their entire lives. Humans use joint attention to coordinate thoughts and behaviours and to cooperate successfully with others. Joint attention is an important factor in learning language, social competence and facilitates social learning. For example, when a teacher admonishes a student to pay attention, in actuality this is really a request to pay attention to what the teacher is attending to. Without the capacity for joint attention, success in many pedagogical contexts would be difficult to achieve [10]. Joint attention is also thought to be important in the development of theory of mind [24], the understanding of one's own and other's minds and the separation between. Theory of mind in turn is a foundational skill for social cognitive functioning, and is related to many aspects of a child's functioning such as social competence, peer acceptance and early success in school [25]. Joint attention is one of the most important skills in social cognition. People who do not have the skills to follow and share attention with others have significant difficulty in relating to other people and sustaining relationships [10]. In fact the lack of these skills in children is often an indicator that a child may have autism spectrum disorder [26].

2.2.2. Joint attention in Human Non-human animal interaction

It is a debated question whether non-human animals can engage in joint attention. Some studies argue that apes do engage in joint attention, while other studies like a work by Siposova and Carpenter [10] argue they do not. Whether animals engage in joint attention with humans is an extension to the question above. In a study by Piotti and Kaminski [27] they sought to find out if dogs would inform an ignorant human about a target that is of interest to the human but not the dog. In this study the claim is made that the dogs that participated established joint attention with the humans.

Regardless of whether non-human animals engage in joint attention, many species are able to follow the gaze or head direction of others. Examples include goats, horses, dolphins, dogs and even tortoises [10]. And humans also respond to the gaze or head direction of non-human animals. In a study by Corneille et al. [28] they showed that the more the perception of the orientation of a dog's head was oriented towards a product, the more valence a human had towards that product. And this relation was linear. Furthermore in a study by Manzone et al. [29] they observed the reaction of humans in regards to gaze shifts of humans, dogs and orangutans. The findings of this study suggest that joint attention can exist between humans and non-human animals. The results indicate that a dog's and orangutan's gaze can shift a human's attention, meaning that humans could engage in joint attention with non-human animals.

2.3. Joint attention between humans and artificial beings

2.3.1. Joint attention in Human-Robot interaction

Collaboration requires goals, knowledge and intentions to be communicated. Gaze can be used to convey these internal states. When collaboration involves a physical environment, interactions can also require the use of referential gaze to reference objects and locations. People are sensitive to robot eye gaze when that gaze is directed at objects or location in the environment. Even when people are not conscious of these gaze cues, they can use the referential gaze cues to for example predict which object to select in object selection games. When the aforementioned mental states are conveyed through non-verbal communication, cooperative tasks are performed faster, errors are detected faster and are handled more effectively than when the communication is purely task based. When subtle gaze behaviour is used to indicate engagement and provide feedback the performance of a human-robot team improves. Having a robot display mutual gaze also improves people's subjective and social evaluation of the robot, furthermore it leads people to see the robot as more intentional and pay more attention to it [14].

Human-human interaction makes use of a variety of perceptual cues, and since robots are becoming increasingly more refined they should be able to exploit these cues for Human-Robot interaction. By having robots make use of human or human-like gaze cues for example, humans can perform better in a human-robot cooperative task. In a study by Boucher et al. [5] they had the participants reach for a cube as indicated by a humanoid robot. There were three experimental setups, one where the robot did not move its head, one where it moved its head but was wearing sunglasses and lastly one where it moved its head and its eyes were visible. The participants performed best if they could see which cube the robot was looking at, and worst when the robot did not move its head at all.

In the study above they showed performance went up when robots purposefully made use of gaze cues, but performance also goes up if a robot ”leaks” gaze cues and humans pick it up sub-consciously as shown in a study by Mutlu et al. [30] Aside from the gaze ”leaks”, they also investigated the effect of the ”humanness” of the robot. In the experiment participants were asked to play a guessing game with a robot where the robot picked an item and the participant had to guess which item was picked out of multiple equidistantly placed items in the table in between them. The participants were allowed to ask yes or no questions to narrow the options down. There were two different robots, one with more abstract features and one very human-like, and two different scenarios, one where the robot did not gaze at the picked item and one where the robot “leaked” a gaze cue towards the picked item. The participants performed significantly better with the gaze cue than without. Furthermore when the game was played with the human-like robot, the gaze cue had a significant positive effect on the performance, but when played with the abstract robot the gaze cue did not significantly affect the performance. This study provided inspiration for the experimental design of this research. Having the participant and the robot settled across each other with the objects of interest in between them allows for easier joint attention, and gazes can be used as cues or hints for the participant.

Joint attention does not only impact performance in HRI, a study by Yonezawa et al. [11] found that humans can “guess” what a robot is interested in by using joint attention and that that eye contact utilised by the robot brings humans favourable feelings towards the robot. In this study they use a stuffed toy animal as the robot for their experiment. In the first experimental setup the robot was placed in between two monitors which played animations. The robot would gaze at both monitors and the participant, but would gaze longer at one of the monitors compared to the other. The subjects had to use the robot's gaze to guess which of the two animations held the robot's interest. In the follow up experiments the set up was almost the same except there were two robots instead of one. In these experiments the robots would display different gaze behaviours. In the follow up experiments the participant had to guess which robot had a more favourable feeling towards the participant using gaze behaviour, joint attention and a combination of the two. It was concluded that joint attention behaviour helped the participants understand the robot's interests, and that the gaze behaviour of the robot is effective at drawing the participants gaze.

2.3.2. Joint attention in Child-Robot interaction

Social robots could be used as therapy tools for children with ASD. Because ASD often comes with a deficit in the understanding and use of social gaze, gaze cues can be a particularly important cue. When interacting with robots, some children with ASD show unprompted social gaze behaviours such as increased eye gaze and shared attention towards robots as opposed to humans [14].

A study by Warren et al. [31] used a robot to help children diagnosed with ASD develop joint attention skills. The participants were told they were going to play a game and then a humanoid robot was used to prompt the participant with joint attention bids. The prompts moved from simple to complex, starting with name and gaze prompts and then moving to include pointing. Participants had 7 seconds to shift attention according to the prompt, if they did so they were rewarded with a clip from preschool TV programs. if the participant did not shift attention according to the prompt within 7 seconds then the system proceeded with the next prompt. This study showed that the performance over time in a basic core communication skill and area of deficit could improve, so children can train joint attention skills by interacting with a robot. A study by Logan et al. [32] showed that the use of robots instead of a stuffed animal for children is viable and feasible, at least in a paediatric setting. In this study they randomly exposed children to one of three interventions, a stuffed teddy bear, an avatar of said teddy bear on a tablet or an interactive social robot teddy bear. Children who interacted with the social robot robot teddy bear reported a more positive effect relative to the other two interventions. Compared to the other interventions, the interactions with the social robot were characterised by greater levels of joy and agreeableness. The two studies mentioned above show that animal-like robots could be used to teach and train joint attention skills in children. Children prefer interacting with a robot over a stuffed animal. Furthermore children with ASD showed improvement in their joint attention skills after interacting multiple times with a robot, albeit a humanoid one. A next step could be to explore if children with ASD also improve if they interact with an animal-like robot.

2.4. How humans view robots matters

2.4.1. Perception & embodiment of robots

There are multiple studies which have investigated the effect of a robot's embodiment on a human's perception of said robot. Studies have shown that humans tend to have more empathy for physical robots compared to virtual [33], that physical robots appear more watchful, more enjoyable [34] and are evaluated more positively in regards to social presence and interaction than a virtual agent [35]. Furthermore a study by Breazeal et al. [36] showed that a physical robot is perceived as a real entity instead of a fictional entity, as is the case with animated characters. The embodiment also has an impact on how humans respond to non-verbal communications of the robot. A study by Abubshait et al. [37] suggests that a physical embodiment elicits a stronger reaction to non-verbal signals such as gaze cues than their virtual counterparts. In another paper by Abubshait and Wiese [13], they claim that the degree to which gaze is followed on an agent depends whether or not they are perceived or believed to have a mind and this is largely affected by if the agent is perceived as a real, intentional being. Furthermore they state that a human appearance and reliable behaviour induce mind perception but in another paper by Wiese et al. [38] it is stated that the effect of appearance only works from a certain point on. An agent needs to be viewed as having mind via their behaviour before a human appearance even matters. Another paper by Eyssel and Pfundmair [39] found that humans can also ascribe mind to animal-like robots.

Although these studies do suggest that physically embodied robots are always better, this is not always the case. One big advantage of virtual agents is that because virtual agents are animated, they can mimic human eye capabilities with greater precision than physical robots can [14].

2.4.2. mental models

So it can be argued that when a robot's capabilities advance, human robot collaboration also improves. But this may not necessarily be the case. When robots display more human social behaviours, humans may develop unrealistic expectations of what a robot can and can not do. A robot may unintentionally cause humans to build an inaccurate mental model of its abilities. A study by Kwon et al. [9] found that both the robot's appearance and behaviour is important as to how humans form these mental models. They found that people tend to generalise the capabilities of humanoid robots more than industrial robots. A study by Walters et al. [40] found that humans tend to perceive Humanoid robots as more intelligent than mechanoid robots. However, if the robot has a short height, they are seen as less conscientious and more neurotic. In a work by Li et al. [41] they found that there is a strong correlation between interaction performance and some matters of preference such as likeability, trust and satisfaction in HRI. Furthermore it was also found that a robot's appearance affects its likeability.

People tend to create a mental model and expectations towards the robot before they even interact with it [42, 43], and improperly designed robots may lead to a disconnect between expectations and reality. This in turn can lead to disappointment and a negative experience.

2.4.3. measuring perception

Perception is how we interpret the world around us, but this is very much an internal process. It is not easy to objectively measure what someone actually perceives, or what they think about what they perceive. Using questionnaires is one of the ways to measure perception, and this will be used in this research project. Two existing questionnaires were considered to be used in this project.

The first questionnaire considered was the godspeed questionnaire [44]. This questionnaire was developed out of a need to be able to compare results from different studies. This questionnaire focuses on five concepts in HRI: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. The questionnaire is not too long, being only 24 questions long and uses semantic differential scales from 1 to 5.

The second questionnaire considered was the Artificial Social Agent (ASA) questionnaire written by Fitrianie et al. [45]. This questionnaire is an answer to the methodology crisis where many scientific studies in social and life sciences are difficult to replicate. This questionnaire makes use of a 7-point Likert scale. In total there are 90 questions and are spread over 19 measurements or subscales. The authors of the ASA questionnaire also created a short version consisting of 24 questions where each question represents a construct or dimension of the full questionnaire.

While the godspeed questionnaire seemed like a good choice, it was not chosen because the ASA questionnaire had more relevant subscales which godspeed lacked, like social presence for example. Social presence is an aspect of this system that is investigated in this project, so the ASA questionnaire is used in this project.

The answer scale of the ASA questionnaire used in this research project is not entirely the same as how the answer scale was originally developed. While both the original scale and the scale used in this research both use a 7-point Likert scale, the labels are a bit different. In the original scale participants gave an answer of -3 to 3, with -3 meaning disagree, 3 meaning agree and 0 meaning neither agree nor disagree. The scale used in this research makes use of the same scoring but different labels.

Work Score	Original	This research
3	agree	strongly agree
2		agree
1		somewhat agree
0	neither agree nor disagree	neither agree nor disagree
-1		somewhat disagree
-2		disagree
-3	disagree	strongly disagree

Table 2.1: Difference in labels between works

3 Experiment Setup

In order to answer the research question and sub research questions as defined in chapter 1, an experiment was needed. For this experiment a game was designed and played by the participants. In this chapter the game will be discussed followed by the experiment methodology. This chapter finishes with the physical setup of the experiment. This experiment was submitted to the TU Delft Human Research Ethics committee and got approved with ID number 3050.

3.1. The Game

3.1.1. Game procedure

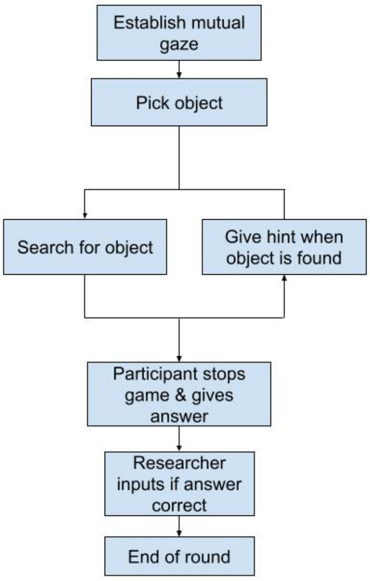

Establish mutual gaze
Pick object
Search for object
Give hint when object is found
Participant stops game & gives answer
Researcher inputs if answer correct
End of round

Figure 3.1: The round procedure

The game was very simple in itself; guess what object the robot is looking for. Figure 3.1 above shows a simple graph of how a round proceeded. At the top of a round the Robot would establish mutual gaze, this was both to have the robot start in a neutral position and a small check for the researcher to see if the system works correctly before a round was started. The robot would then randomly select one of the nine objects as the object it would search for. The robot would search the nine objects and would give a non-verbal hint whenever it ”found” the desired object. It would search by looking at each object one by one in one of four predetermined search patterns. Whenever the robot reached the end of a search pattern, it would follow the same search pattern in reverse back to the start. Once the robot was looking at the same spot it started at, it would consider the pattern complete and select a new one. See the figures 3.2 through 3.5 below for the search patterns. Whenever it looked at an object it was not looking for, it would move on to the next object in the search pattern. When it looked at the object it was looking for, it would give a non-verbal hint before moving on to the next object. The robot would continue this cycle of searching and hinting until the participant stopped the game and gave an answer. The participant could only give one answer per round, and afterwards they would only be told if the answer was correct or not. No hint would be given as to how wrong the answer was or what the correct answer should have been. Aside from inputting the answers into the system, the researcher was otherwise in no way involved with the workings of the game, i.e. there was no wizard of OZ going on. Aside from checking the answers and inputting them into the system, the system performed the game fully autonomously. The non-verbal hint of the robot used joint attention with different levels of jointness; the levels of jointness as described in section 2.1. Each level would make use of more gaze behaviour and should have given more information as to what the intended object was. Both the participant and the robot could see all the objects, the robot however initially only looked at the objects and thus did not know if the participant was looking at the same object. The participant knew the robot was looking at the objects but was not sure which one as this was the very task of the game. This means that both parties were aware of the objects but uncertain on which object the other's attention was focused on, so the first level of joint attention as described in 2.1 was achieved. Each game would last five rounds, and a total of three games would be played, one for each level of joint attention.

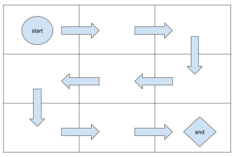


Figure 3.2: Pattern 1

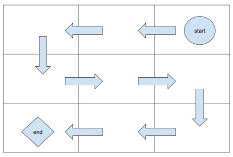


Figure 3.3: Pattern 2

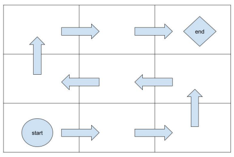


Figure 3.4: Pattern 3

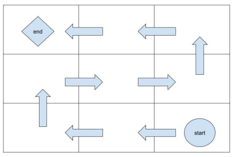


Figure 3.5: Pattern 4

3.1.2. Why a game

The goal of the experiment was to see if different levels of joint attention had an effect on the opinion and perception participants had towards the robot and their performance in the game. The goal of the game was to do this by using the subconscious behaviour of joint attention whilst also preventing the participant from realising it was about joint attention. This is why participants had to go in as "blind" as possible. If they already knew the experiment was about joint attention, they might know what to look for and potentially skew the results. The reason it was designed as a game is because it was deemed easier by the researcher to hide the actual goal from the participants in a game than in a task. Aside for the reasons from the perspective of the experimental design, games may be a fitting interaction between children with ASD and robots anyhow; in order to see if the interaction quality between a child with ASD and a robot changes over time, a study by van Otterdijk et al. [46] investigated the long term effects of child robot game interactions. They measured non-verbal behaviour such as gaze and arm/hand movements towards the robot, the game and other humans. Their analysis showed no significant decrease in the attention and the engagement towards the game and the robot. They did find however that the attention and engagement towards parents increased. van Otterdijk et al. concluded that this sustained attention and engagement towards the robot is due the personalisation of the games to meet the specific needs of the participants.

Although the study was performed with adults, according to Banks et al. [47] how a person perceives the interaction affects their judgement towards the robot in different ways. If a participant perceived the interaction as a play interaction, it fostered moral trust in the robot and shifted the participants attitude over time. If the interaction was perceived as a task interaction it fostered functional trust in the robot. A game could be perceived as both a play interaction and a task interaction, meaning it could foster both moral and functional trust in the robot.

3.2. Methodology

3.2.1. Type of experiment

This experiment had a completely randomised within-subjects design with one factor; the level of jointness. All participants would play three games, one for each joint attention level, but the order in which they played these games was counterbalanced. The experiment used a blind setup, the participants do not know which level of joint attention the robot was employing during the experiment. Participants should not have access to this knowledge before or during the experiment as it was unwanted that the participants make inferences about this level. The researcher performing the experiment was aware which level of joint attention the robot used. In the debrief after the experiment participants would be told their order and what the levels mean.

3.2.2. Participants

In a work by Cohen [48] they define effect sizes for social and behavioural sciences as 0.1 for small, 0.25 for medium and 0.4 as large. For a repeated measures ANOVA with a power of 80%, three measurements and significance level of 0.05, 12 participants are required to detect a large effect size. For a medium effect size 28 participants are needed, and this number increases to 163 if the effect size to be detected is small. The other input parameters, correlation among repeated measures and nonsphericity correction, were kept at their default value of 0.5 and 1 respectively. G*Power [49] was used for these calculations. According to a video made by Dr. Björn Walther [50], and a module written by Dr Mark Williamson [51], the sample size for a Friedman test would follow the same calculation as the repeated measure ANOVA, but the result should be increased by 15%. This would mean that for the Friedman test 14 participants are needed to detect a large effect size, 33 participants to detect a medium effect size and 188 participants to detect a small effect size. In order to properly counterbalance the experiment, there should be an equal number of participants for each order of games that could be played. Since there were three games to be played, there were six distinct orderings of games. This meant that the amount of participants should be divisible by 6, as this would allow for an even spread among the ordering. Due to time constraints, the number of participants was chosen to be 30. This number allowed for an even spread among the orderings and would exceed the minimum for detecting both large and medium effect sizes with a power of 80% in parametric tests and large effect sizes in non-parametric tests. If time would have allowed it, another six participants would have been recruited to make sure medium effect sizes for non-parametric tests could be detected with 80% power and to keep the number divisible by 6. In the end, the extra participants were not recruited keeping the number of participants at 30. All the participants were older than 18, did not need glasses to see sharp between 1 and 2 metres and were recruited from social circles. Participants were offered a snack and something to drink, but most declined the snack.

3.2.3. Hypotheses

Joint attention has an impact on a human's mental model: in the study by Yonezawa et al. they saw an increase in favourable feeling towards the robot if joint attention is used [11]. Furthermore, joint attention makes use of gaze behaviour and in a study by Boucher et al. [5] they found that humans performed better the more clear the gaze cues were. So the hypothesis for the main research question was as follows:

> **Main hypothesis: An increase in the robot's joint attention level leads to a more positive interaction between a robot and a human.**

The main research question was divided into three sub research questions. And for each sub research question hypothesis was formed. As mentioned above, in a study by Boucher et al. [5] they saw that more clear gaze behaviour improved task performance. Furthermore, a study by Mutlu et al. [30] showed that humans are receptive to the gaze cues of a robot. In this study they found that when the robot made use of gaze cues the performance of participants was significantly better than when the robot did not use gaze cues. So the hypothesis for SRQ 1 was:

Sub hypothesis 1: An increase in the robot's joint attention level leads to an increase in the human's task performance.

Mental model includes subjective measures such as whether a participant likes the robot or trusts the robot to perform its job. Having a robot display mutual gaze improves people's subjective and social evaluation of the robot. It leads people to see the robot as more intentional and pay more attention to it [14]. When robots use social gazes such as mutual gaze instead of neutral non-social gazes, humans feel more engaged with the robot, enhancing the social interaction with the robot [52]. This leads to the hypothesis for SRQ 2:

Sub hypothesis 2: An increase in the robot's joint attention level leads to a positive effect on the human's mental model of the robot.

What is meant with a positive effect on the mental model is something like whether a participant likes the robot more after the interaction or they have more trust that the robot knows what it is doing. The effect could also be negative where the participant likes the robot less for example.

The degree to which gaze is followed on an agent depends whether or not they are perceived or believed to have their own mind. If a robot is perceived as having its own mind, humans respond better to non-verbal communication. This perception of a robot as a being with its own mind is largely affected by if the agent is perceived as a real, intentional being [13]. This might be because as mentioned above, people pay more attention to a robot if they see the robot as more intentional. And people see the robot as more intentional when it displays mutual gaze behaviour [14]. So the hypothesis for SRQ 3 was:

Sub hypothesis 3: An increase in the robot's joint attention level leads to an increase in the human's perception of the robot as its own entity with its own mind.

3.2.4. Measures

In order to perform an experiment the variables necessary need to be defined. For all sub research questions, the same independent variable was used: joint attention level. This variable had three levels; the three different levels of joint attention as defined in chapter 2 on page 5. For each sub research question, different primary outcome measures were used. For SRQ 1, the primary outcome variables were accuracy, time needed, and hints needed, all per game. These three objective variables were used to measure task performance and are continuous dependent variables. These variables were automatically recorded by the experiment system, for the system can easily keep track of start and stop times, pauses, how many hints were needed and whether an answer was correct or not. It already has access to this data during the experiment, it only needed to be recorded. SRQ 2 and 3 used subjective measures, therefore a questionnaire was needed aside from the experiment. SRQ 2 and 3 were about opinions so these needed to be asked instead of measured. In order to take these questionnaires, Qualtrics [53] was used. The questionnaire used was the ASA questionnaire described in section 2.4.3. Since the questionnaire needed to be taken three times, the entire ASA questionnaire was too large, while the short version was too shallow. Fortunately each subscale could be used on its own as long as all the questions within a subscale were used. In order to keep the questionnaire as short as possible to prevent questionnaire fatigue, the amount of subscales was kept to a minimum. As stated in chapter 2, how humans view a robot matters, so the selection of the subscales focused on this. Ultimately four subscales were chosen, these subscales were: the agent's likeability, the agent's attentiveness, the agent's social presence and the agent's intentionality. For SRQ 2, the primary outcome measure was the total score calculated from all four subscales. This measure was used as a continuous dependent variable. For SRQ 3 the primary outcome measure was the score calculated from only one subscale; the agent's social presence. This score was used as the continuous dependent variable for SRQ 3. In order to explore and help explain the data, 10 secondary outcome measures were collected. One variable called scenario was collected to keep track if data was from the first, second or third game. The other nine measures are categorised as demographic variables:

1. Highest education followed.
2. Highest education completed.
3. Experience with robots.
4. Experience with animals.
5. What pets they own/have owned.
6. What animal they think the robot looks like.
7. How familiar they are with joint attention.
8. How well they know how joint attention works.
9. Whether they think the robot ”points” with their eyes or nose.

3.2.5. Experiment procedure

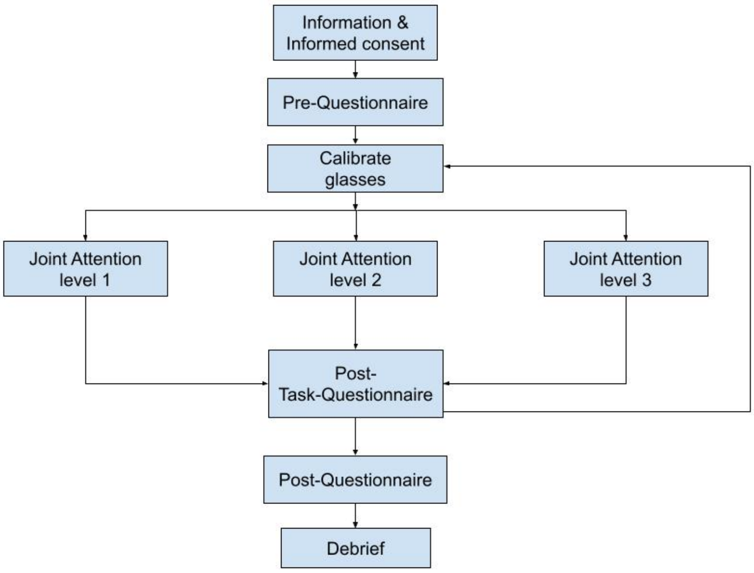


Figure 3.6: The experiment procedure

Figure 3.6 above shows how an experiment proceeded. First the informed consent form would be given which would explain what data would be collected and why, all the while trying to prevent priming the participant. The researcher would briefly explain the experiment itself. The only explanation the participants would get is the amount of games and rounds played, the fact that the robot was searching for a random object each round, that the hint is non-verbal, that the robot would continue searching and hinting until the participant stopped the game and that the participant could stop the game at any time and should not wait for a designated moment to give their answer.

After the brief explanation a pre-questionnaire was read out loud to the participants while the researcher recorded the answers on their own device. The pre-questionnaire collected the first six demographic variables as defined in section 3.2.4. Because of the way Qualtrics works, it was easier to make pre, post-task and post questionnaires into a single large questionnaire divided by sections. This way all answers had to be recorded on the same device. In the pilot studies it became clear rather quickly that it was easier to have the researcher read the survey questions out loud and record the answers on their own device than to have the participant do it themselves. The reason was that while none of the participants required glasses to see sharp at 2m, some did require reading glasses. So instead of having the participants switch out the pupil labs glasses for reading glasses and having to completely recalibrate the pupil labs glasses each round, participants kept the pupil labs glasses on and the questions were read to them. This way not every round required a recalibration, this was only done where necessary.

After the pre-questionnaire, the pupil labs glasses were calibrated. The calibration of the glasses only happened when necessary such as after the glasses were removed or put on for the first time. Following the calibration a game would be played with a joint attention level that was random for the participant. Each participant played three games, one for each level of the independent variable. After every game a post-task questionnaire was read out loud to the participant. This post-task questionnaire collected the scenario variable and the primary outcome variables for SRQ 2 and 3. Once all the games were played and all the post-task questionnaires were filled in, the post-questionnaire would be read out loud to the participant. This questionnaire consisted of only three questions and collected the last three demographic variables defined in section 3.2.4. These three variables were only collected after all the tasks in order to prevent priming the participant. The final part of the procedure was the debrief. Now that priming no longer mattered, the goal of the game and experiment were explained and participants were free to ask any questions. Before a participant left they were asked not to discuss the game and experiment with others until the experiment was done.

The experiment lasted on average around 40 minutes in total, including briefing before the experiment, calibrating and the debrief afterwards. A single round took on average a little less than 3 minutes, but the time it took for the questionnaires was not timed.

3.3. Setup of the Experiment

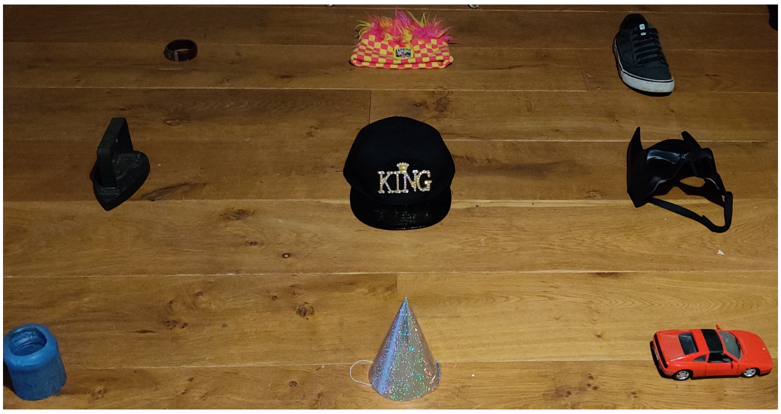


Figure 3.7: The nine objects the robot can search for

3.3.1. The game

The participant was seated in a chair 2 metres across from the robot. The robot was also seated in a chair, this was because of the way the robot was designed. The robot is incapable of looking down, so in order for it to do so it was seated on a chair in a tilted position by using books to prop the back upwards a bit. This way it would at least look like the robot could look down. In between the participant and the robot were nine objects laid out in a 3x3 grid. Each object was placed exactly 50 cm away from its neighbouring object. While this may seem trivial, by keeping the distances exact, it is much easier to calibrate the robot. The reason nine objects were chosen instead of four is that four would be too little. The game would have been too easy with only four objects. The opposite issue would arise with 16 objects. The game was deemed difficult enough with a 3x3 grid. A 4x4 grid would have made the game harder and that was deemed unnecessary for this experiment, but might be interesting to use in further research. The nine objects were chosen to be as unambiguous as possible, so that when the participant would give their answer, there would be no doubt as to which object the participant is referring to. The nine objects were: a shoe, a beanie, a bracelet, an old iron, a cap, a mask, a model car, a party hat and a candle. They can be found in figure 3.7 above:

3.3.2. Experiment environment

Because of an issue with the pupil labs glasses, which will be discussed in more depth in chapter 4, the lighting in the room had to be controlled. No bright lights could be in front of the robot or the experiment would not work. So all light from the windows had to be blocked out with curtains and blankets. This however caused the room to be too dark for the robot detector to work. So in order to prevent this, a white sheet with a bright light was placed behind the robot. This way the robot would not be affected by the light, and the robot detector would be capable of seeing and detecting the robot. Furthermore, the researcher had to be seated in the room, both to keep an eye on the system and intervene if things went wrong, and to judge if the answers of the participants were correct or not. Figure 3.8 below represents the spatial configuration of the experiment environment. The blue square in this figure represents a wall with a rack which formed a little nook where the researcher could be seated next to the participant. Here the researcher would be just out of the peripheral vision of the participant, so while the participant was paying attention to the game, the researcher would not give accidental hints with involuntary facial expressions. This little nook also caused the researcher to be in the shade of the only lamp of the room. This enabled the researcher to have an overview of the game, while also preventing the robot from seeing and reacting to the researcher.

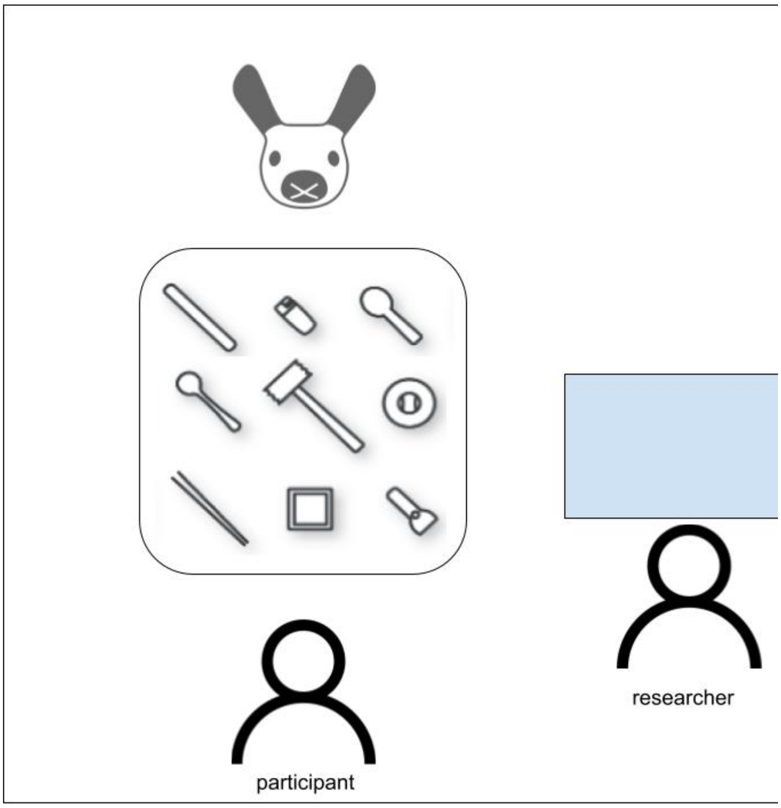


Figure 3.8: The spatial configuration of the experiment

3.4. System requirements

In order to perform the experiment as explained above, a system was needed. A pre-existing system did not exist so one needed to be developed. Now that what the experiment aimed to do and how to do it is settled, a list of requirements for the system could be drawn from it:

	Requirements
R1	The system must be able to collect and save the necessary data.
R1.1	The system must be able to track the performance measures.
R1.2	The researcher must be able to input if a participant's guess is correct or wrong.
R2	The participant must be able to stop the system at any time so they can give their answer.
R3	The system must be able to detect when joint attention is established.
R3.1	The system must be able to detect if the robot looks at a human.
R3.2	The system must be able to detect if a human looks at the robot.
R4	The robot must be able to perform three distinct levels of joint attention.
R5	The robot must be able to "look" at nine distinct places on the floor.

Now that the requirements are established, a system fulfilling those requirements can be designed and implemented. This design and implementation will be discussed in the next chapter.

4 Design and implementation

In order to perform the experiments, a system was needed. In this chapter the design and implementation of the system is discussed. First up is the main design principle with its design choices. Afterwards the hardware used is described briefly since it is not the main focus but still a relevant section. The next section is about the software of this project. This section is divided into two parts: first is software tools and applications developed outside this project. This software was not actively developed but was used in this project as is. The second part of the software section is developed software. This software was as the name implies actively developed during this project and the code can be found at: `https://gitlab.ewi.tudelft.nl/in5000/ii/joint-attention-with-an-animal-like-robot`

4.1. Design choices

4.1.1. Main design principle

The main design principle of this system was reusability. There are a few reasons as to why this was the main design principle. The first reason was that as mentioned in chapter 2 the appearance of a robot has influence on the mental model of a human and a human's perception of the robot as an entity with its own mind. The mental model and perception in turn influence a human's response to the non-verbal communication of the robot. Appearance was not something that was investigated in this research project, but it could be an interesting factor to investigate in the future. So if this system could easily repeat this experiment with a different robot that would greatly facilitate the ease of which future research could expand on this research. The second reason was that the end goal of this system is to be able to help children with ASD, but not every child needs the exact same help. By making this system easily reusable with the possibility of adjusting and tweaking the system to a child's needs, the help these children receive could be personalised. And according to a work by van Otterdijk et al. [46], personalisation works in favour of keeping a child's attention and engagement toward both the robot and the game.

4.1.2. Design choices

Outside of accounting for the needs for the experiment, the system was also developed with a few design choices in mind. These design choices were not necessary to perform the experiment, but with exception of the first, these design choices were added with the idea that this system could be reused in the future.

Programming language

The first design choice was the programming language used: Python. The software development kit of the robot offers the choice of using Python, C or C++, but the reason Python was chosen was twofold. Most of the documentation uses Python as an example development language. While it is mentioned on consequential robotics' website however that C or C++ users will find that all the same things are possible. The other reason was that Python was the preferred language of the researcher developing this system, and they have much more experience in programming in Python than in C or C++.

Modularity

The second design choice comes partially forth from the chosen programming language and partially because it makes reusability easier: modularity. With Python as the chosen programming language, it is easy to make every method its own file, and every file its own module. This system did not go to that extreme, but the general idea was that the system should be developed with the future in mind. For example one idea was to reproduce this experiment, but with different robots. This way the impact of robot design could be measured. But not every robot has the same degrees of freedom or even uses the same ROS messages to communicate. So the parts in this system which were robot specific or behaviour specific were designed to be easily swapped out. This way a researcher only needs to develop a few parts instead of an entire system. The swapping of these parts does not even need to be hard-coded, the system was designed to be able to choose which robot and behaviour profile to use on start up.

Separability

The third design choice was separability. Some of the parts of this system were on the heavier side with regards to computing power. For example both the robot detector and the face follower required a bit of processing power to do what they were designed to do. So some of the heavier parts were designed to be able to run on different PC's and work together over a network. These parts already used ROS messages to communicate, so it was a small step to make it possible for them to communicate over a network. A sufficiently powerful PC is capable of running all parts of this system locally at once, as this was how the experiment was performed. But not everyone has access to a sufficiently powerful PC, or it may be easier to source multiple less powerful PC's.

User-friendliness of setting up/ modifying experiments

The fourth design choice was to make setting up/modifying the experiment user-friendly. This was not necessary per se, the experiment could have been hard-coded. But in the same vein as the second design choice; by making it easy to set up or modify experiments it is also easy for other researchers to use this system to make their own experiments. An example of this is the movement builder tool.

By including this in the system it was easier to create or adjust movements and search patterns. Furthermore it would make it much easier for future researchers to make their own movements and behaviours. And while it was not used in this experiment as the robot used the same behaviour profile for everyone, in the future these profiles could be personalised and this would work well in keeping children's attention and engagement toward the robot and the game [46]. Timings could be adjusted according to someone's preferences or to test what works best. Furthermore different movements could be created for different games. The system was designed in such a way that on startup it gives the choice of which behaviour profile should be used. Because of time constraints the movement builder only works for robots who make use of the same ROS messages for movement as the MiRo-E, but because the controller was used to actually make the robot move, it should not be difficult to adapt the movement builder to a new controller.

Another example of this design choice was to make the detection model maker user friendly. This part had to be made anyway for the robot detector. By making it easy to train new models it also made it easy to experiment with experimental set ups. Because of this, the issue with the infrared light which could have delayed the experiment significantly was solved in less than two days. Another advantage of designing the model maker this way was that it was easy to create a model for a new robot. It takes around 90 images of the robot and 2 to 3 hours on a good PC to train a model. It takes a researcher less than half a day to set up the robot detector with a new robot, this includes manually taking the images, manually labelling the images and letting a good PC train the model for 10 epochs.

4.2. The Hardware

This system made use of two pieces of hardware, the MiRo-E robot and the pupil labs glasses. The section below contains a brief description of the hardware and figure 4.1 contains an overview of how the hardware is connected.

4.2.1. MiRo-E robot

The MiRo-E is an animal-like robot developed by Consequential Robotics [54] in association with universities and educationalists. They designed the MiRo-E to be animal-like so that people interact with and respond to the MiRo-E with a different set of expectations as opposed to when the robot would be human-like. The MiRo-E has a wide-ranging suite of sensors, these include stereo vision location hearing, ultrasonic ranging, light level sensors, infrared cliff sensors, tactile sensors on the body and head, and interceptive sensors such as twin accelerometers. It is also capable of proprioception and sensing temperature [55]. The MiRo-E was an excellent choice for this research project because it is designed to be used by children, even below the age of seven, as both an educational tool and a tool for cognitive therapy [56]. Furthermore the MiRo-E is also designed with research in mind. Two areas of research the designers explicitly mention that the MiRo-E is suited for is research into HRI and robot assisted therapy [57].

4.2.2. Pupil Labs Glasses

Eye tracking is the process of measuring eye positions, pupil positions and eye movements. This data can then be used to determine where someone is looking or what they are looking at. Eye tracking technology usually has multiple cameras: one for each eye and one to look at the world. An often used method is to use infrared light to detect the difference in reflection between the pupil and the cornea, also known as pupil centre corneal reflection [58]. Pupil Core is an open source software suite and a wearable eye tracking headset. The headset consists of lens-less glasses with three cameras mounted to it. Two of these cameras, the so-called eye cameras, are mounted on the bottom of the glasses and face towards the wearer's eyes. These cameras track the position of the pupils, allowing the glasses to track where the wearer is looking. A third camera is mounted on the brow of the glasses and faces forward. This so-called world camera allows the glasses to track what the wearer is looking at. These three cameras together allow the glasses to perform gaze tracking on the wearer in real time [59].

4.3. Software used as is

4.3.1. Software tools

ROS

ROS is an open-source middleware tailored for robotics. And while it stands for Robot Operating System, it is not a true operating system. Although it does offer functionalities expected from an operating system. Within ROS, there's a ”graph” of processes that form a network, which can be spread across different machines. These processes communicate through ROS's infrastructure, supporting various communication styles like synchronous RPC, asynchronous data streaming, and data storage on a Parameter Server. It's important to note that ROS isn't real-time, but it can be combined with real-time code if needed [60]. The reason it is important is because participants were actively playing a game with the robot where time was measured. Therefore the ROS messages between parts of the system need to be transported in real-time in order to perform the experiment. Fortunately ROS is capable of this with its streaming of data over topics [60].

MDK

The MDK or MIRO Developer Kit, is a software package which a developer needs to install on their workstation. It includes everything required for working with a MiRo-E robot [61]. The MDK can be used in either Python, C or C++ and makes active use of ROS. This software package also includes examples on how to make the MiRo-E do some simple movements but these examples are only available in Python. These examples served as a great basis to figure out how to make the MiRo-E do what it needed to do for the experiment.

4.3.2. Software applications

Pupil capture

The Pupil Core headset has its own dedicated capture software called Pupil Capture. Pupil Capture is part of the open source software suite developed by Pupil labs [59] mentioned in section 4.2.2. Pupil Capture is responsible for processing the video feeds from both the world camera and the eye cameras. The video streams of the eye cameras are used to identify and monitor the wearer's pupils. It then uses this pupil data to track their gaze, and in combination with the data of the world camera detects and follows the wearer's gaze in their surroundings [62].

Pupil publisher

This piece of software has been developed in-house by Wouter Pasman of the TU Delft but it is not publicly available. This software connects to pupil capture using the network capabilities already available in the pupil capture software. It then converts the data to ROS messages and publishes it to ROS. This software acts as the bridge between the pupil capture software and ROS. It is a standalone piece of Java, so no installation of ROS is necessary. And because it communicates over a network, this software can also be run on a separate machine than the pupil capture software.

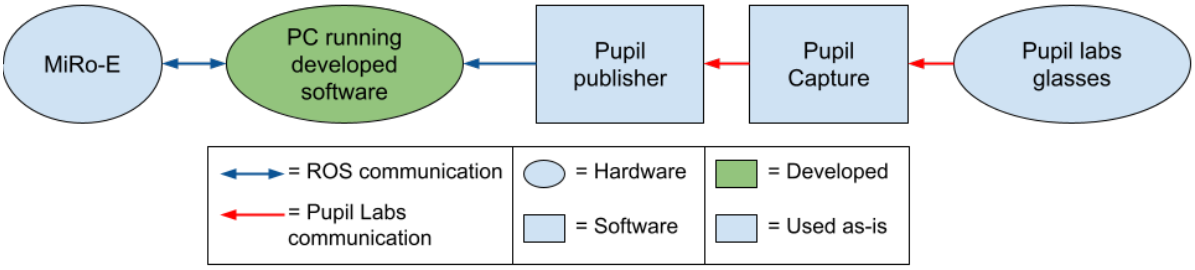


Figure 4.1: Overview of how the hardware is connected

4.4. Developed software

The software described below was developed explicitly for this research project. It was made so that the experiment described in chapter 3 could be performed. Furthermore it was designed so that others could either replicate this experiment, or perform their own. To facilitate this further, some tools were developed to create and adjust robot behaviour, and to easily train the robot detecting neural network necessary for mutual gaze detection on a new robot or in a new environment. Figure 4.2 below contains an overview of how the software is connected.

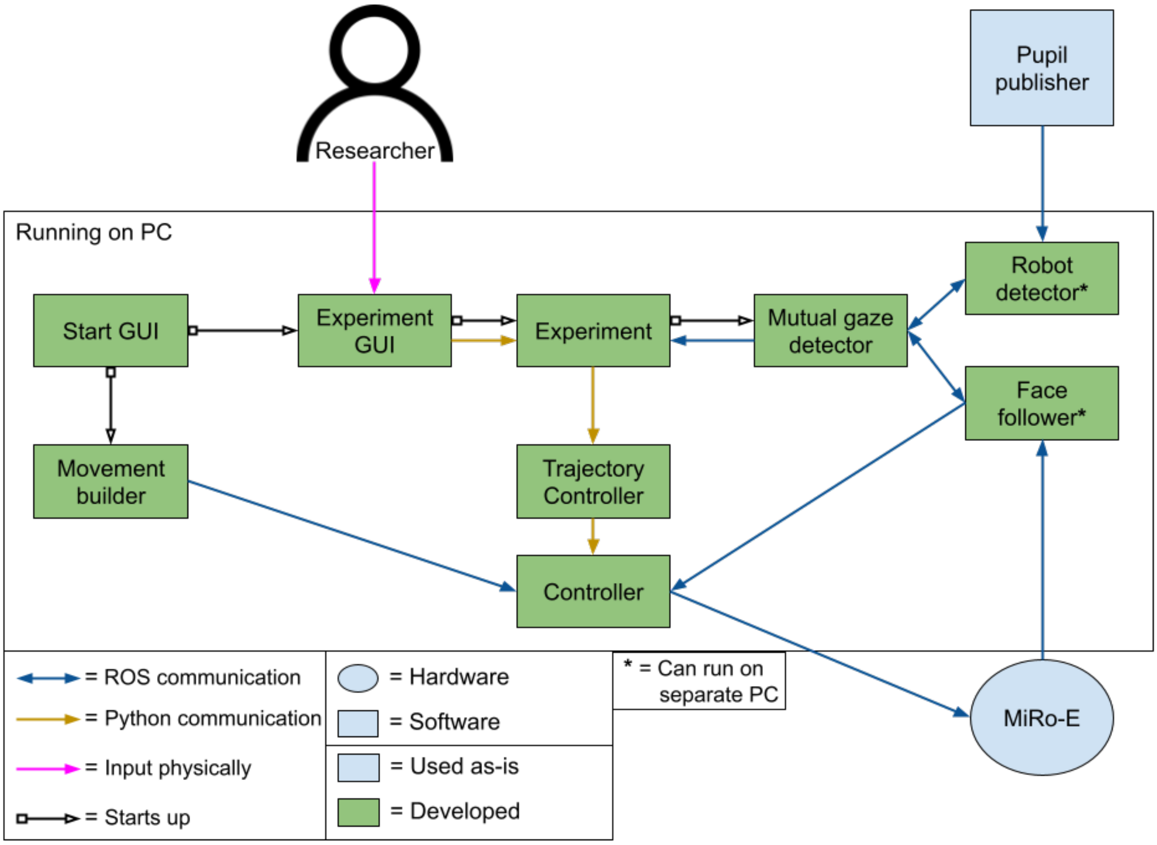


Figure 4.2: Overview of how the software is connected

4.4.1. Experiment

Experiment.py was the script responsible for the experiments, and consists of two classes; the experiment class was responsible for actually running the experiment. It determined the search pattern the MiRo-E used, the object MiRo-E was looking for, and was responsible for collecting and saving the data. The experiment script kept track of how long the participant was taking per round and per try, whether they were correct in their guess, how many hints they needed per try and of course how many tries they had guessed correctly. From this data other data like accuracy could be calculated. The experiment class was also responsible for when the hints should be performed and which level of joint attention to use. How long the robot spends hinting is dependent on how fast it could establish mutual gaze, meaning that hints did not necessarily take the same amount of time every time a hint was performed. Therefore the experiment class was also responsible for recording this hint time so that the time measure could be adjusted. This way a more accurate representation of the time a participant needed for a game could be used for the analysis. The experiment class also spawned mutual gaze detector threads when necessary.

An unused functionality of the experiment class is determining when to use idle behaviour such as acting bored or checking whether the participant is still paying attention. The intention behind this functionality was that by introducing movement that was not a hint, it would desensitise participants to the sound the robot made while moving, specifically the change of sound when movements changed. In the end this behaviour was not used in the experiments themselves for they would add unnecessary extra variables, but the code was left in case another researcher in the future would like to make use of it.

The experiment_gui class was responsible for starting the experiment with the correct value for joint attention level and for pausing and resuming the experiment. It also allowed the researcher to input whether the participant was wrong or correct in their guess.

Start_gui.py

Start_gui.py was an auxiliary script. This script was the starting point of this system. It allowed the researcher to set which robot was used and which behaviour profile were used. From this script either experiment.py or the movement builder was started. It had no functionality other than this, but it provided the starting point for the interactable part of the system.

Controller.py

This part of the system was actually intended to be swapped out whenever necessary. Because not every robot has the same degrees of freedom or even the same limbs for that matter. And because not every robot communicates using the same ROS messages, the controller was made to be robot specific. It takes the more general commands from the experiment and translates them to ROS messages the specific robot can interpret. Which controller is used was checked and loaded in at runtime, so switching controllers is almost the same as just restarting the system. One advantage of keeping the robot specific code in a single file was that its easier whenever a new robot is used with this system. Because only a new controller would have to be written instead of needing an entire system.

Trajectory_controller.py

Originally the robot would make very mechanical movements, whenever it would receive a new coordinate to move to it would move with a constant speed. This movement did not look natural in the slightest, so an effort was made to improve this. Hard-coding fluid movement was not the right choice. The intention was that researchers should be able to make and adjust movements on their own without having to manually make the movements more fluid. This would take up too much time every time a new behaviour was made. So to do this mathematically, the company behind the robot, consequential robotics, was contacted. They came up with a simple formula:

$$curr\_pos = start\_pos + (end\_pos - start\_pos) * move\_progress$$

The current joint position would then be sent to the robot at 50 Hz. While this was a great start, it was not smooth enough, it needed more coordinates to move through. So this formula was used as a first step, the system would simulate creating and moving to new coordinates or points using the formula until the current point was the same as the end point. It would then use the linspace function of numpy to evenly make more points in between the calculated points. Finally it would use all the points calculated in the previous step and use numpy's interp function to interpolate even more points. By using all three steps, the movement looked much smoother and less mechanical. It achieved this by having more points in the beginning and end of the movement, making it look like it speeds up and slows down. Much of this code is actually in the controller, and only called from the trajectory controller. An attempt was made to move it all to the trajectory controller, but for unknown reasons this caused the code to break and stop functioning.

4.4.2. Mutual Gaze Detector

The mutual gaze detector was a very important part of the system. Without this the entire underlying principle of joint attention would be very difficult to test.

The mutual gaze detector consists of three parts: the mutual gaze detector itself, the robot detector and the face follower. The robot detector uses the model produced by the five steps of the detection model maker combined with the data from the pupil labs glasses to detect if the participant is looking at the robot.

The face follower actually works in two ways. The original way is by using the cameras on the robot, this script will try to detect the participant's face and then move the robot's head to keep the face in the centre of its view. Unfortunately some cameras can actually detect infrared light, so if someone wears the pupil labs glasses the face will be a white blur preventing the script from detecting faces. This was solved by having the robot track the brightest point it can detect instead. This works as well as tracking faces but does require the experimental setup to take this into account, e.g. no lights in the robots field of view can be brighter than the pupil labs glasses. Figures 4.3 and 4.4 below illustrate these two "modes" of face tracking. The picture in figure 4.3 was taken more up close so the face detection is more visible for a human observer, but both modes work at the same distance. The face detection part of this system is based on a tutorial written by Sabbir Ahmed [63].

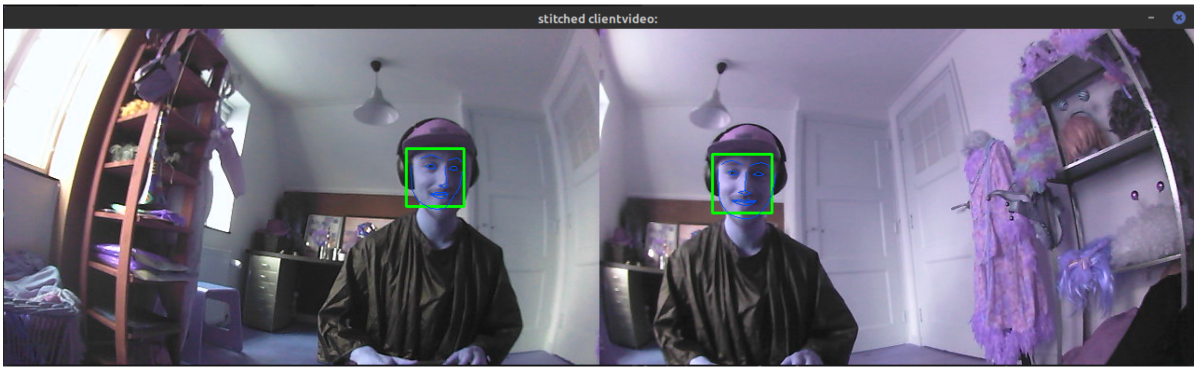


Figure 4.3: When the Face follower tracks faces

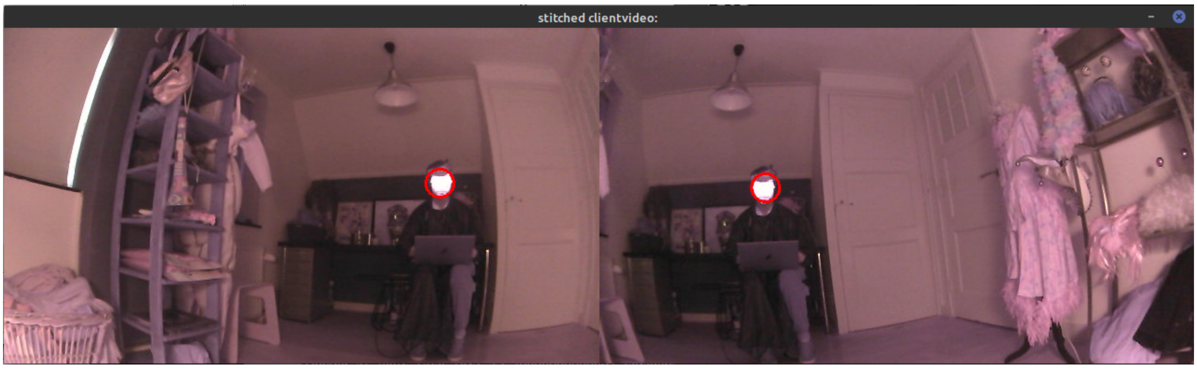


Figure 4.4: When the Face follower tracks lights

This controlled lightning in the room does cause the experimental set up to be relatively dark. This made it difficult for the next part of the system, the robot detector, to detect the robot. Because of the ease of training a new model, it is possible to train a network in these low light conditions and this will work without issue if trained correctly. An example of this can be seen in the figure 4.5 below, the blue box is the boundary of what the script identifies as MiRo-E. The red box is the pupil data of the Pupil Labs glasses. Whenever one of the corners of the red box is within the blue boundary, the script will interpret this as the participant looking at the robot. That only a single corner needs to be within the blue boundary instead of the entire square is a deliberate choice because the pupil data is not 100% stable and accurate. This causes the red square to jump around. By making the detection ”easier” this way, the robot detection and therefore the mutual gaze detection go smoother.

The mutual gaze detector itself was more of a manager of the other two parts. It facilitated communication between the face follower and the robot detector and let the rest of the system know if mutual gaze was happening. By running all the parts of the mutual gaze detector as separate scripts, two goals were achieved; 1. because of the way Python is written, threads spawned within a single script do not run parallel but sequential. So too many threads slows everything down way too much. By running them as separate scripts, but still as a thread, they could run parallel. And 2. the second goal achieved with this is that it worked towards the third design choice. This is because the scripts communicate via ROS so they can be run on separate machines altogether. A strong enough PC can run all scripts on their own, but if such a machine is not available the load can be divided amongst multiple machines.

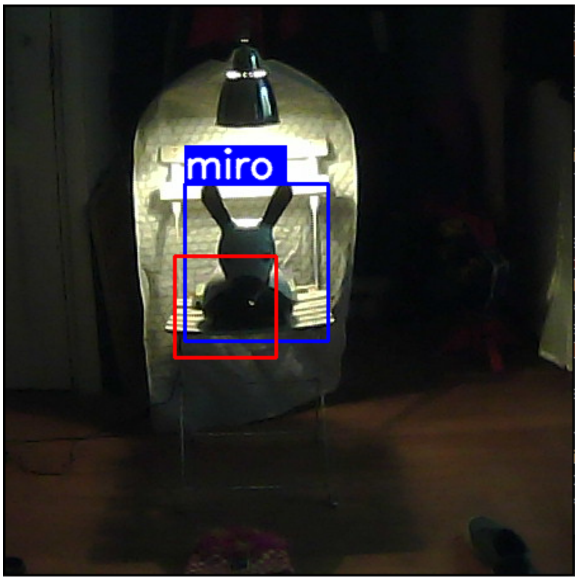


Figure 4.5: What the RobotDetector ”sees”

4.4.3. Movement builder

At first, all the movements the robot made were hard-coded and tested by hand. The movement builder was made to make this easier. By making a GUI where movements could be saved, adjusted and queued, anyone could make a set of behaviours using the movement builder. The GUI has sliders for the pitch, lift and yaw positions for the joint states of the MiRo-E, so by moving the sliders, the robot also moves. This way, a researcher could for example set out nine objects in a grid and move the sliders until the robot looked at an object, save the exact position and repeat for the other eight objects. Because it is possible to load pre-existing movements and overwrite them with tweaks, it made adjusting them and recalibrating after setting up the objects a breeze. Because the movement builder has access to the cameras of the robot, it can also be used to calibrate its or the participants position to an extent. For example by being able to see the robot's point of view, the researcher could quickly detect if the participant is in view of the robot and stays in view when the robot moves. Because the movement builder has this functionality, this can all be tested quickly without having to start up the entire system. With the exception of once in the beginning when the experiment environment was set up, this functionality was not used much in this research project. This is because the experiment did not need to be set up and broken down multiple times, however for experiments where this is the case, it could speed up the setup of said experiments.

4.4.4. Detector model maker

In order for this system to detect mutual gaze, it needed to be capable of detecting if a human is looking at the robot. But in order to do that, the system first needed to be able to detect the robot itself. Since the robot is an object, an existing technique called object detection would work perfectly. Object detection is a technique within computer vision that involves identifying instances of objects in images or videos. Algorithms for object detection commonly rely on machine learning or deep learning techniques to produce meaningful results. Humans can recognize and locate objects of interest in a video or image within a matter of moments. With object detection it is possible to replicate this using a computer.

This part of the system was designed to be as easy to use as possible. The reason for this is that by making it easy for researchers to create new models, it would be easy to repeat experiments with different robots. This in turn facilitates the reusability of the system. Another point of note is that it would also make it easy to change environments or setups. This point is noteworthy because of one limitation of the robot detector; the robot detector works best if trained in the same environment and setup as how the experiment will be performed, when this is not the case the robot detector still works but it is noticeably worse. The neural network used here is not of self-made design. Because the system is largely a prototype, designing a neural network from scratch was deemed out of scope. The system makes use of TensorFlow's implementation of the convolutional network VGG16. This network is designed by Simonyan and Zisserman [64]. A free course published on youtube by Nicholas Renotte [65] was used as a tutorial to train the network. This allowed for a significantly shorter development time of this part of the system, and thus allowed for more focus on other parts of the system. From start to finish it only takes a few hours to train the network on a new robot. This allows for relatively quick adjustments to either the environment or the robot itself, which in turn allow for more flexibility. There are five steps involved in this process, and as such this part of the system is divided into five parts.

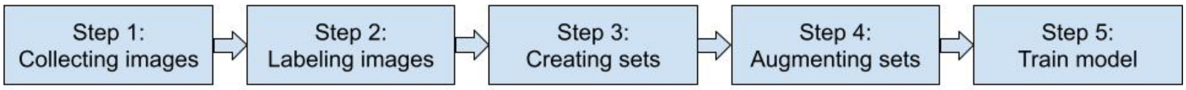


Figure 4.6: Pipeline of the model maker

Step 1: collecting the images

The first step is collecting images on which to train the network. It is not necessary to use the same camera to take these images as the camera that is used by the system to recognise the robot, but it is highly recommended. As such the system allows for the user to choose which camera they want to use. When using the same camera and lighting set up as in the experiment, around 90 images are enough. These 90 images can be taken in less than 15 minutes. The images are immediately saved using an uuid in their name, so the user does not need to name every image, and does not need to worry about name conflicts or duplicates. Images are automatically saved in a folder hierarchy suitable for the next steps, no user input is needed here except specifying the name of the root folder when this script is run.

Step 2: labelling the images

For this step no script was written, but a library called labelme [66] was used instead. By using labelme, the pictures taken in step 1 can be hand annotated using custom labels. If 90 images were taken, using labelme would take around 15 min. The labels created by labelme are json files that use the same name as the images, so it is easy to identify which label belongs to which image.

Step 3: making the sets

This part of the system is quite simple. This script separates the images and labels into training, testing and validating sets. It takes 70% of the images for the training set, and the remaining images are divided between the test and validating sets. The script keeps the labels and images together and generates the necessary file structure expected by the next step if it doesn't exist yet. The user only needs to specify a root folder name, which needs to be the same as in the first step.

Step 4: augmenting the sets

90 images is not a lot to train a neural network on. The reason it is enough for this system is this step. It uses a Python library called albumentations to augment the sets. It takes the images and labels and randomly crops them, randomly flips them horizontally, vertically or both, it randomises the brightness contrast, randomises the gamma and does some RGB shifts. It does this 60 times per image so that instead of 90 images, the system can train on 5400 images, and these are all different because of the randomisation.

Step 5: training the model

This step takes all the augmented images and labels from the train set and trains a model. If 90 images were taken in step 1, this will take a decent PC about 2 hours. So in about 3 hours, a new model for a new robot or experimental setup can be created

Helper functions

The image label combiner:
The images and the corresponding labels have the same name, but they are saved in different folders to keep everything ordered. The image label combiner combines these into a single dictionary for future use. It also resizes images to make them a suitable size for the VGG16 neural network.
Model builder:
Step 5 is responsible for training the model with the right parameters, the model builder is responsible for actually making the model and determining how it is trained. In model builder the loss functions are set which step 5 will make use of.

5 Analysis

In this chapter the analysis will be discussed. First the data preparation will be explained after which the main analysis is described in depth. In the main analysis all the dependent variables will be analysed and conclusions for the hypothesis as defined in chapter 3 will be drawn. The research questions will not be answered here, this is done in chapter 6. Following the main analysis some exploratory analysis will be discussed. This exploratory analysis section is for analyses which did not directly answer the sub research questions, but were still of value to the main analysis.

5.1. Pre-processing

This analysis was done in Rstudio using the language R version 4.3.1. But first some pre-processing has been done in Python. The pre-processing served a few purposes, one of which was to calculate the variable adjusted time. Adjusted time is the total time of a game minus the hint time of a game, so this variable actually represents how much time a participant needed irrelevant of how long the robot took with its hints. Hint time is the time it took for the robot to convey its hint. The reason the system kept track of hint time was because part of the hint was the robot trying to establish mutual gaze. This meant that the hint time was variable and not the same for everyone. This was partly because the establishing of mutual gaze was dependent on the participant, i.e. how fast the participant looked at the robot influenced the speed at which mutual gaze could be established. And partly because the data from the pupil labs glasses was not 100% accurate, there was some variability as to when the system would detect the participant looking at the robot. Even if a participant were to look at the robot in the exact same way with the exact same timings, there would still be a little variation in the mutual gaze detection speeds. Because of this the hint time was kept track of, and adjusted time was used as a performance measure. The second function of the pre-processors was to calculate the scores of the questionnaire. Each answer in the questionnaire was associated with a score ranging from -3 to 3. The pre-processor added all the scores within a subscale together separated by scenario. So each participant would have 12 sub total scores: four for each scenario. The sub-scores with the same scenario were added together to get the total score and the score for the social presence subscale was saved separately. In the end a participant would have six scores associated with them, for each scenario a total and a social score. Because of the way the survey was set up, it was easier to calculate these scores afterwards and to separate these scores by scenario. Another function of the pre-processors was to remove all the unnecessary data from the Qualtrics survey data. Aside from the data actually needed it also included data such as the IP address from where the survey was taken. This data was irrelevant and only caused clutter, so it was removed for the analysis. The last purpose of the pre-processor was to merge the survey data with the performance data so that all the data necessary for the statistical analysis would be contained in a single file. While the code for the analysis can be found on the gitlab page mentioned in chapter 4, the data used in this analysis is not publicly available. To get access to this data, see the gitlab page.

5.2. Main analysis

In this initial data analysis the assumptions for performing ANOVA are checked before any analysis is performed. First the residuals of the measures were checked to see if the data is normally distributed. A linear mixed model was fitted on the data and the residuals were plotted in both a density and a QQ plot. Complimentary to visual confirmation, a Shapiro-Wilk test was also performed on the data. Homogeneity of variance was checked with Levene's test, and where Sphericity needed to be checked Mauchly's test of Sphericity was used. Sphericity needed to be checked because the data was from a repeated measures designed experiment. More details about the initial data analysis can be found in Appendix B. Because ANOVA is robust against violation of the assumption that the residuals are normally distributed [67], ANOVA was used in the analysis if there was only indication for the violation of the normality assumption and there were no outliers. If there was indication for the violation of two or more assumptions, or there was indication that the normality assumption was violated together with the presence of outliers, ANOVA was not used. Because the experiment was a repeated measures design and therefore not all data is independent, other parametric tests would also not be used. So instead the non-parametric Friedman test was used. This test was chosen since it is a non-parametric alternative to repeated measures ANOVA [68]. And because Friedman is rank based, it is also resistant to outliers. The effect size of the ANOVA tests was calculated using Cohen's formula as published in [69] with the generalised eta squared obtained from the ANOVA test. For Friedman the effect size was calculated with: W = X2/N(K-1). Where W is the Kendall's W value, X2 is the Friedman test statistic value, N is the sample size, and k is the number of measurements per subject. By using G*Power [49], these effect sizes were then used to calculate the statistical power if statistically significant differences were found.

5.2.1. SRQ1; the performance measures

SH 1: An increase in the robot's joint attention level leads to an increase in the human's task performance.
To test if this hypothesis is supported, three measures were analysed: time, hints and accuracy. These measures were analysed individually before a conclusion about the hypothesis was formed.
Time:
When looking at figure 5.1 below, it looks like time needed went down as the robot made use of higher levels of joint attention. Not only does the violin itself become smaller, The red dot in the plot representing the mean also goes down. This suggests that joint attention had a positive effect on the time a participant needed to complete a game.

Figures B.1 and B.2 show the residuals visualised in both a density and QQ plot, the Shapiro-Wilk test showed that the distribution of the residuals of time departed significantly from normality: W = 0.80904, p = 1.938e-09. Next was to check the homogeneity of variance using Levene's Test. Levene's Test showed that there was no significant difference in variance with F(2,87) = 1.0094, p = 0.3687. A visualisation of the variances can be found in figure B.3. The next step would be to check sphericity but this was not needed since on top of the evidence for violation of the normality assumption, time had multiple outliers as seen in figure 5.7. So Friedman was used instead of ANOVA, with Nemenyi as post-hoc analysis.

Friedman showed that there was a significant difference between joint attention levels with X2(2) = 11.267, p = 0.003577. Post-hoc analysis using the Nemenyi test between joint attention levels revealed statistically significant differences between level 1 and 3(q= 4.747,p = 0.022762). The effect size of this test was 0.188, meaning this was a small effect according to Cohen's size index [48]. Using G*Power it was calculated that the statistical power of this test is 58.7%.

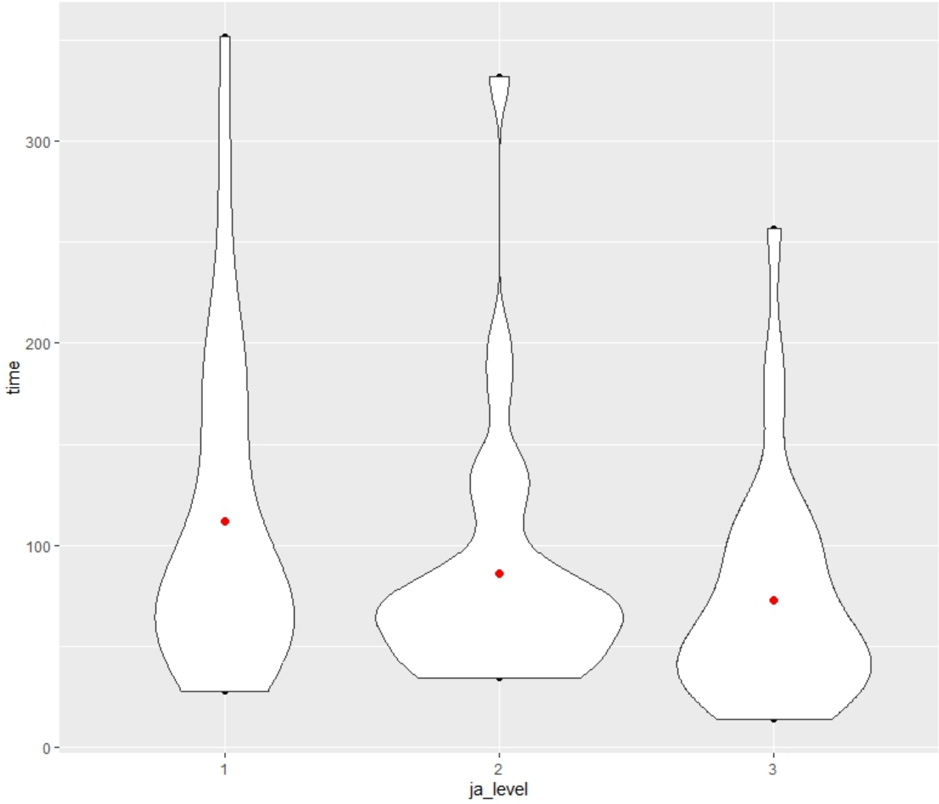


Figure 5.1: Mean time by joint attention level

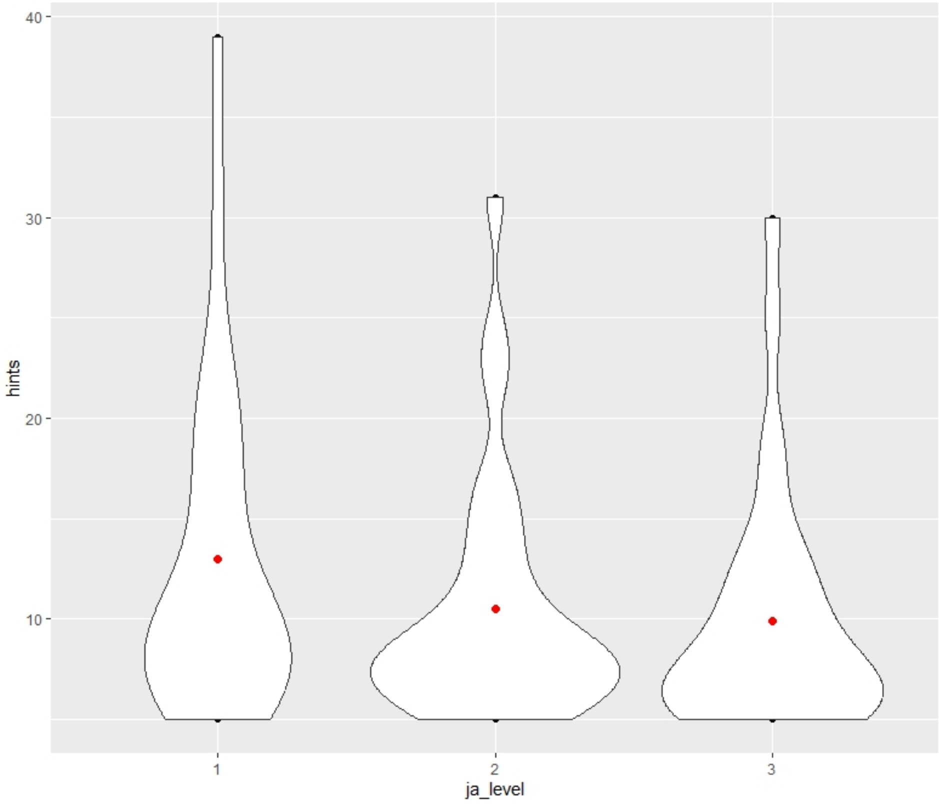


Figure 5.2: Mean hints by joint attention level

Hints:
Figure 5.2 above shows that when the robot made use of higher levels of joint attention, the amount of hints needed went down, especially between levels 1 and 2 of joint attention. Like with the time measure, the red dot representing the mean goes down. This suggests that joint attention had a positive effect on the amount of hints a participant needed to complete a game.

Like with the time measure, a Shapiro-Wilk test was performed aside from visually checking the distribution as shown in figures B.4 and B.5. The test showed that the distribution of the residuals of hints also departed significantly from normality: W = 0.80786, p = 1.785e-09. The next assumption to check was the homogeneity of variance using Levene's Test. Levene's Test showed that there was no significant difference in variance with F(2,87) = 0.9919, p = 0.375. A visualisation of the variances can be found in figure B.6. Sphericity also did not need to be checked here just like with time since hints had outliers as seen in figure 5.8 on top of the evidence that the normality assumption was violated. So like with the time measure, Friedman was used with Nemenyi as post-hoc analysis.

Friedman showed that there was a significant difference between joint attention levels. With X2(2) = 7.4579, p = 0.02402. Post-hoc analysis using the Nemenyi test between joint attention levels revealed statistically significant differences between levels 1 and 3(q = 3.56, p = 0.031712). The effect size of this test was 0.124, meaning this was a small effect according to Cohen's size index [48]. Using G*Power it was calculated that the statistical power of this test is 28.6%.
Accuracy:
Unlike the measures above, the heights of the violins in figure 5.3 do not change with different levels of joint attention. This means that for every level of joint attention, at least one participant got a 100% correct and one participant got 0% correct. What does change however is the mean represented by the red dot. The figure shows that for higher levels of joint attention the mean goes up, meaning more participants scored high. This suggests that higher levels of joint attention did allow for better scores, and thus had a positive effect on accuracy.

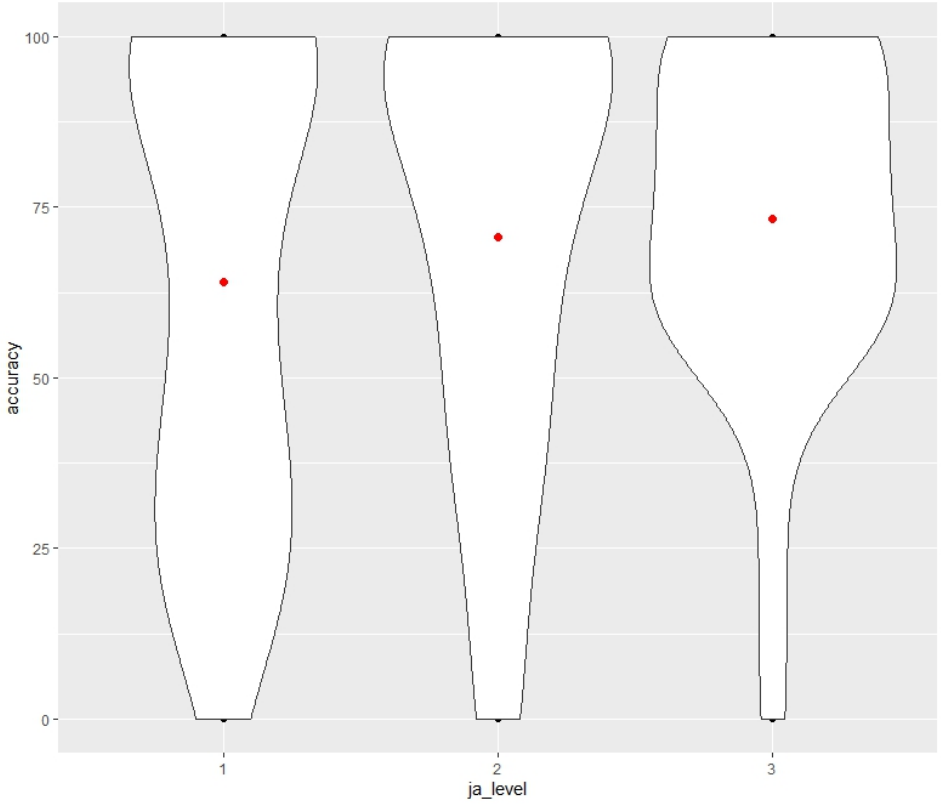


Figure 5.3: Mean accuracy by joint attention level

As with the measures above, a Shapiro-Wilk test was also performed next to plotting density and QQ plots of the residuals of the accuracy measure. The plots are shown in figures B.7 and B.8. The Shapiro-Wilk test showed that the distribution for the residuals of accuracy were not normally distributed: W = 0.91181, p = 1.423e-05. Like the previous measures the homogeneity of variance was checked using Levene's Test. Levene's Test showed that there was a significant difference in variance with F(2,87) = 3.2444, p = 0.04376. A visualisation of the variances can be found in figure B.9. This means that there was evidence that the data for accuracy violated two assumptions, so ANOVA was not used here. Instead, Friedman was used here with Nemenyi as post-hoc analysis.

Friedman showed that there was no significant difference between joint attention level with X2(2) = 0.96296, p = 0.6179. No post-hoc analysis is performed.

Result:

The results showed that for time and hints joint attention had a significant and positive effect on the measures. The post-hoc tests revealed this to be between levels one and three of the joint attention level for both measures. So when the robot used a higher level of joint attention it led to higher task performance, at least in these aspects of task performance. For accuracy no significant difference in the means was found. Since two out of three measures for task performance did increase when the level of joint attention increased and for the third there is no significant effect positive or negative was found, it was concluded that the hypothesis stated above is supported.

5.2.2. SRQ2; the total score

SH 2: An increase in the robot's joint attention level leads to a positive effect on the human's mental model of the robot.

To test if this hypothesis is supported, the total score measure was analysed. The violin plot as shown in figure 5.4 shows that the means did not change much with joint attention level. This suggests that joint attention level may not have had much of an effect on total score.

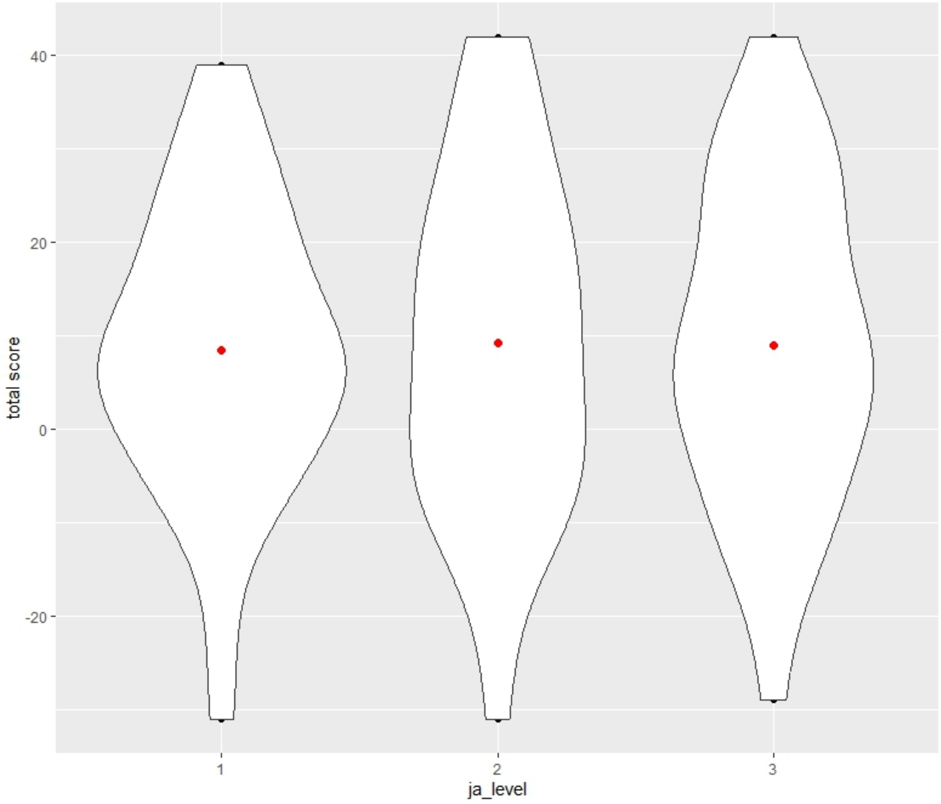


Figure 5.4: Mean total score by joint attention level

Like the analysis of the performance measures, a density and QQ plot were plotted with the residuals of total score and a Shapiro-Wilk test was performed. The plots can be found in figures B.10 and B.11 and the test showed that the difference with a normal distribution was non-significant: W = 0.98558, p = 0.4239. Since there was no indication that the normality assumption is violated, the next assumption to test was the homogeneity of variance using Levene's test. Levene's test showed that the variances did not differ significantly, with F(2,87) = 0.793, p = 0.4557. A visualisation of the variances can be found in figure B.12. Sphericity was tested using Mauchly's Test of Sphericity, and this test showed that the data had sphericity with F(2,87) = 0.13, p = 0.879. No evidence was found for any violation of the necessary assumptions, so repeated measures ANOVA was used. ANOVA showed that for total score there were no significant differences in the mean between joint attention levels: F(2,87) = 0.01695, p = 0.9832.

Result:
The results show that for total score no significant difference in the group means was found. it was concluded that there was no evidence found to support nor any indication to reject the hypothesis stated above.

5.2.3. SRQ3; the social score

SH 3: An increase in the robot's joint attention level leads to an increase in the human's perception of the robot as its own entity with its own mind.
First an initial data analysis was performed just like with the sub research questions above. Like with total score, the violin plot as shown in figure 5.5 shows that the means did not change much with joint attention level. This suggests that social score may not have been affected much by joint attention level. The residuals of the linear mixed model were plotted in a density plot and a QQ plot and a Shapiro-Wilk test was performed. The residuals of the data for the social score were not normally distributed. This is visible in both the density plot and the QQ plot found in figures B.13 and B.14

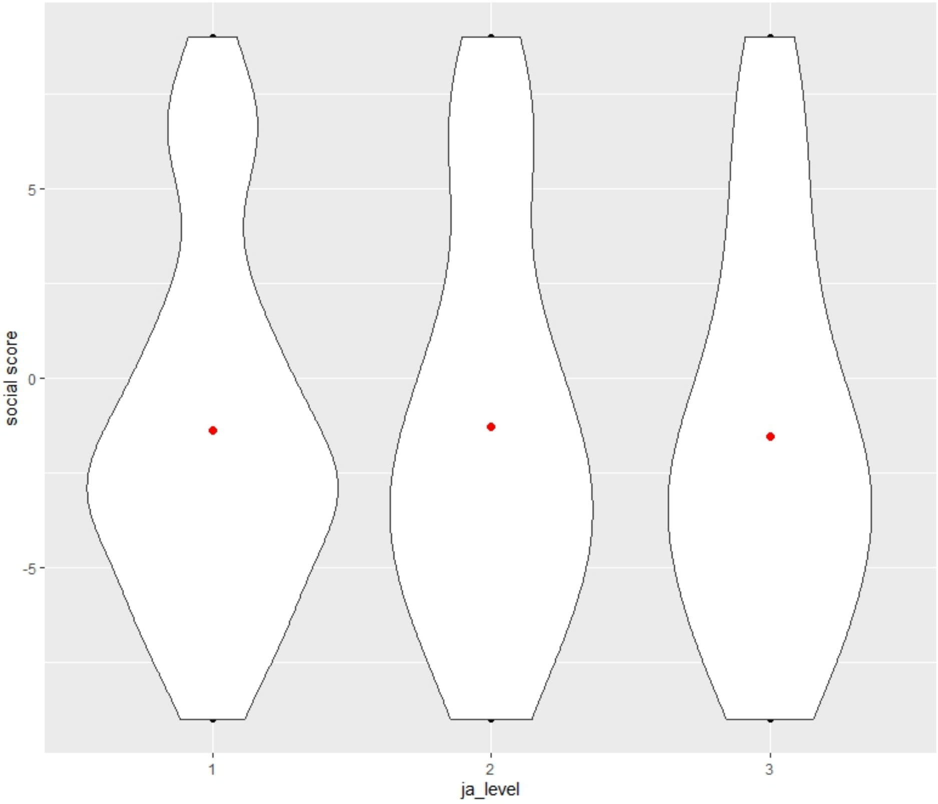


Figure 5.5: Mean social score by joint attention level

Furthermore the Shapiro-Wilk test also showed the distribution had a significant departure from normality: W = 0.94159, p = 0.0005251. So it is safe to say that there was evidence to support that the normality assumption was violated. But just like with SRQ 1, the other assumptions are still tested since ANOVA is quite robust when it comes to normality violations. The next assumption to test was the homogeneity of variance using Levene's test. Levene's test showed that there was no significant difference in variances, with F(2,87) = 0.1262, p = 0.8816. A visualisation of the variances can be found in figure B.15. There was no indication that the assumptions for ANOVA were violated, but like SRQ 2, Sphericity needed to be tested before repeated measures ANOVA could be used. Mauchly's Test of Sphericity showed that the data had sphericity with F(2,87) = 0.13, p = 0.879.

All the assumptions necessary were tested and there was only evidence that the normality assumption was violated. Furthermore, social score had no outliers as seen in figure 5.11, so repeated measures ANOVA was used. This test showed that for social score there were no significant differences in the mean between joint attention levels: F(2,87) = 0.01766, p = 0.9825.

Result:

Like with total score, the results show that for social score no significant difference in the group means was found. it is concluded that there was no evidence found to support nor any indication to reject the hypothesis stated above.

5.3. Exploratory analysis

5.3.1. Data exploration

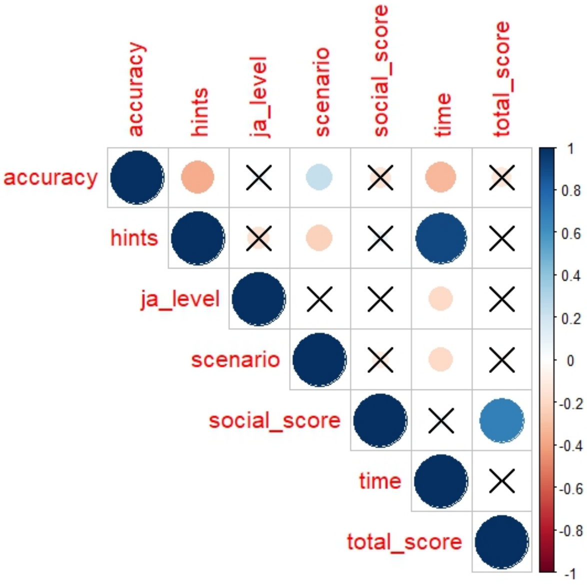


Figure 5.6: Correlation plot

Correlations

The above figure 5.6 shows a correlation plot, the crosses in the plot represent non-significance. The Appendix D also contains a table with the exact numbers. All the non-demographic variables were compared to every other non-demographic variable. Also included in this analysis were the variables joint attention level and scenario. Some relations were not surprising, for example hints and time. Other relations were more surprising, like accuracy and total and social score, they were negatively related. It was interesting to see that joint attention level was only significantly correlated to time, but scenario had a significant correlation to all the performance measures, hinting that scenario had an effect on the performance measures. Total and social score both did not have a significant correlation with joint attention level and scenario, suggesting that these measures were very stable, and not affected much by the independent variables.

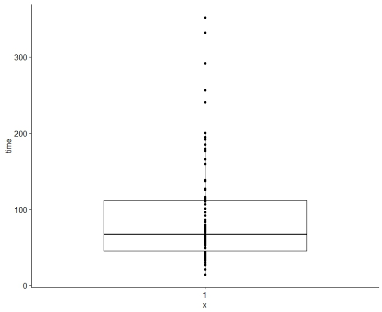


Figure 5.7: Box plot of time

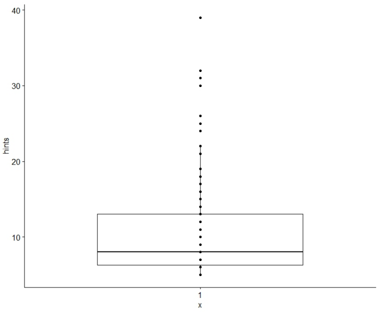


Figure 5.8: Box plot of hints

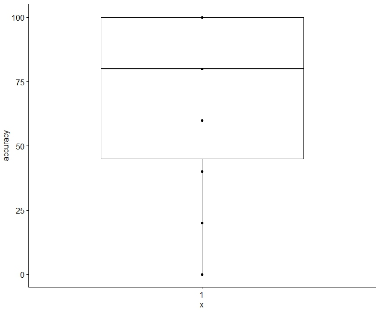


Figure 5.9: Box plot of accuracy

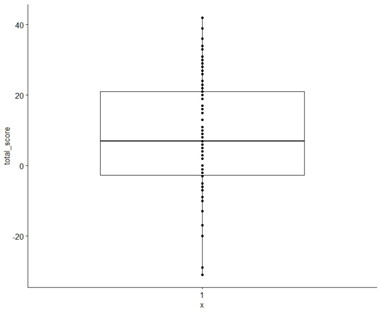


Figure 5.10: Box plot of total score

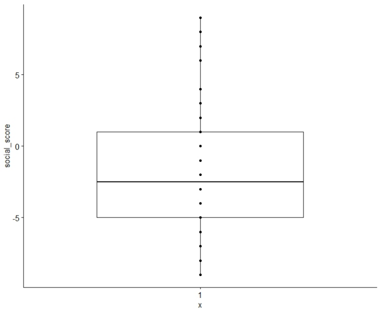


Figure 5.11: Box plot of social score

Outliers

Outliers are not considered assumptions that can not be violated in order to use ANOVA like homogeneity of variance, but outliers can have a significant impact on ANOVA tests. They can skew the results and affect the validity of the test. Outliers were both checked visually with boxplots and the function in the rstatix package in R called: identify_outliers() [70]. The function calculates four quartiles and the interquartile range IQR(IQR = Q3 - Q1), a value is considered an outlier if it is below Q1 - 1.5 * IQR or above Q3 * 1.5IQR. Using the rstatix function the exact outliers could be identified, the exact outliers are also listed in Appendix D. The boxplots in figures 5.7 through 5.11 above show that only time and hints have outliers. Because there was a limited amount of data and combined with the fact that the residuals of these measures were also not normally distributed, Friedman was used in the analysis of these measures instead of removing these outliers. Friedman is resistant to outliers because it is rank based.

5.3.2. Investigating ordering effects

Scenario was an internal variable to keep track of the order of which level of jointness was used in which game. Scenario 1 was always the first game, 2 always the second and 3 always the last. So if a data point had a value of 2 for scenario and 3 for level, then this meant that the second game played made use of jointness level 3 for the hint. The variable scenario was created to synchronise the survey data to the performance data, this way it was easy to keep track which survey answers belonged to which game. Scenario, like joint attention, had three levels: 1, 2 and 3. As shown in figure 5.6 scenario had a significant correlation with the performance measures. Using scenario instead of joint attention as the grouping factor when making violin plots revealed some interesting data. As shown in the figures 5.12 through 5.16 below, it looks like scenario had an effect on the performance measures. Contrary to joint attention level, scenario looks to also have had an impact on total and social score.

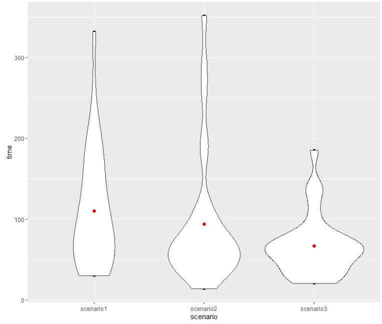


Figure 5.12: Mean time by scenario

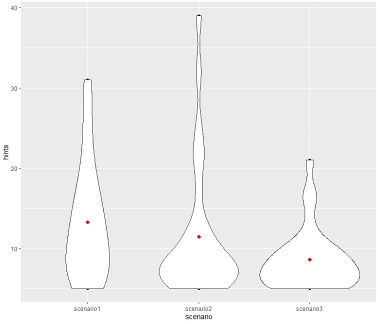


Figure 5.13: Mean hints by scenario

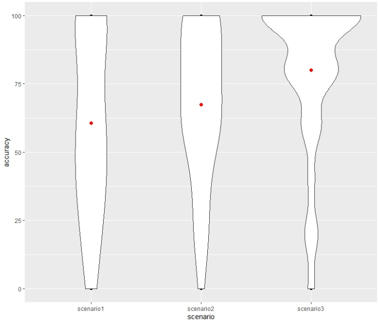


Figure 5.14: Mean accuracy by scenario

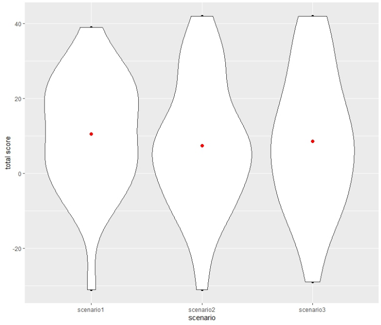


Figure 5.15: Mean total score by scenario

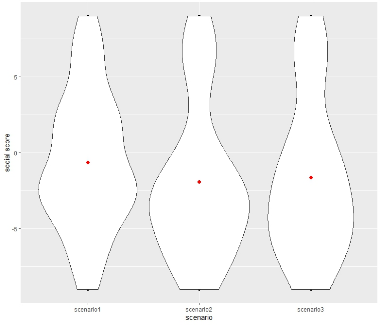


Figure 5.16: Mean social score by scenario

In this analysis the same steps were performed as in the main analysis, but then using scenario instead of joint attention level as the independent variable. Since scenario could be viewed as a measurement of a point in time, if this analysis would reveal significant results this could be interpreted as an ordering effect. Below is a shortened analysis containing just the results of the tests, the full analysis can be found in Appendix C.

Performance measures

Just like in the main analysis, there was indication that the data for the performance measures violated at least two of the assumptions: the assumptions of normality and sphericity. So the non-parametric Friedman test was used. For both the variables accuracy and hints, Friedman showed that there was a significant difference between scenarios. With X2(2) = 11.062, p = 0.003963 for accuracy and X2(2) = 8.7664, p = 0.01249 for hints. There was no significant difference between scenarios for time with X2(2) = 2.6, p = 0.2752. Post-hoc analysis using the Nemenyi test between scenarios revealed statistically significant differences between scenario 1 and 3(q = 3.651, p = 0.026564) for accuracy and between scenario 1 and 3(q = 3.925, p = 0.015219) for hints.

Total score measure

Just like in the main analysis, there was no evidence to support that the assumptions necessary were violated, so repeated measures ANOVA was used. However, this test showed that for total score there were no significant differences in the means between scenarios: F(2,87) = 0.2502, p = 0.7792

Social score measure

Unlike in the main analysis and like the performance measures of the exploratory analysis, there was indication that two of the assumptions necessary for repeated measures ANOVA were violated: normality and sphericity. so Friedman was used instead of ANOVA. This test showed that for social score there were significant differences in the mean between scenarios: X2(2) = 9.9277, p = 0.006986. Post-hoc analysis using Nemenyi showed that there was a statistically significant difference between scenario 1 and 2(q = 3.651, p = 0.026564).

5.3.3. No multivariate analysis using demographic variables

Originally, the demographic variables were collected to help explore the results of the main analysis. After all the data had been collected it became clear that this may not be worthwhile. The demographic variables were very heavily skewed as shown in Appendix E, and a model selection using AICc revealed the most applicable model to be a model with only one demographic variable: experience with animals. This variable had five levels and only three of these levels received a response. Of these three levels with a response, one received over 75% of the responses. Since there were nine demographic variables in total, and almost all of them were heavily skewed it was deemed that this analysis would not be worthwhile and therefore this exploratory analysis is omitted. The table 5.1 below shows the demographic variables with the mode and range of each variable. For more details about the divide of responses of the demographic variables, see Appendix E.

Variable	Mode	Range
Education following	None	None mavo havo vwo mbo hbo wo bachelor wo master phd
Education completed	hbo	None mavo havo vwo mbo hbo wo bachelor wo master phd
Experience working with robots	None at all	None at all A little A moderate amount A lot A great deal
Experience working with animals	None at all	None at all A little A moderate amount A lot A great deal
Pets owned	cats	open ended question
What animal does MiRo resemble	rabbit	open ended question
Familiarity with joint attention	Not familiar at all	Not familiar at all Slightly familiar Moderately familiar Very familiar Extremely familiar
Knowledge of how joint attention works	Not well at all	Not well at all Slightly well Moderately well Very well Extremely well
Points with eye(s) or nose	Eye(s)	Nose Eye(s)

Table 5.1: The demographic variables with their Modes and Range

6 Discussion

In this chapter the results of analysis and limitations will be discussed and conclusions will be drawn. First the results of the main analysis and the exploratory analysis will be covered. In the main analysis section some expectations of this research will be discussed, followed by a discussion about the ordering effects found in the exploratory analysis. The discussion ends with some interesting findings about the demographic variables. Next some limitations will be discussed and what effect they might have had on the results. Finally, this chapter ends with the conclusions drawn in this research project and the research question and its sub research questions will be answered.

6.1. Discussion

6.1.1. Main analysis

While the results of SRQ 1 were within expectation, hence the hypothesis, it was surprising to find that level 3 did better than level 2 as much as it did. The original expectation was that level 2 would do better than level 1 as it would display more information with its hint behaviour (and level 3 would do better than level 1 for the same reason). But the difference between levels 2 and 3 was expected to be smaller, namely that for most people level 2 would be enough to get the hint and level 3 would not add much to it. Interestingly, a few participants actually did worse with level 3 than level 2. This could mean that for those participants, level 3 was too much movement and instead of giving the participant a clearer hint, it caused them to be confused.

The result of SRQ 2 was an unexpected outcome of this project. The expectation was that as people performed better, they would form a more positive mental model, i.e. they would like the robot more, or trust it more. But the results were inconclusive, meaning that this is not necessarily the case. When looking at the correlations, the performance measures had a very small correlation and in some cases it was even negative. Although all these correlations were not significant, people tend to form the mental model and expectations towards the robot before they even interact with it [42, 43] and it seems that the games played in this experiment were not enough to sway someone once they formed their mental model.

Just like SRQ 2, the result of SRQ 3 went against expectations. It was expected that the higher the level of joint attention the robot used, the more it would be perceived as a social entity. The reason for this expectation was that joint attention is a social subconscious behaviour, so by displaying more of this behaviour humans would view the robot as more social. It was also expected that as the robot uses higher levels of joint attention, the behaviour would be seen as more reliable since performance also goes up. This in turn would then induce mind perception in the participants [13]. But at least in the context of this experiment, this was not the case.

6.1.2. Analysis of the ordering effect

Because scenario can also be viewed as a measurement of a point in time, by using scenario as the independent variable instead of joint attention it could be investigated if a learning effect could apply here. In the analysis of the performance measures it showed that for accuracy and hints scenario had a significant effect. This could mean that participants got better at the game as they played it more, which would not be illogical. The results show however that scenario had no significant effect on time. This is interesting as time and hints were highly correlated, and since the data table in Appendix D suggests scenario did indeed have an effect. It could be that the Friedman test gave a false negative, it could also be because the variance of time is much higher than the variance of hints as shown in the table 6.1 below.

And unlike when joint attention was used as the independent variable, with scenario as independent variable a significant effect was found for the social score. The effect was negative however, meaning that as participants played games they thought that the robot had less of a social presence than when they started. When looking at the data in table D.10 and figure 5.16, the social score in the first scenario was higher than the other scenarios while the social scores of scenarios 2 and 3 were more even. The analysis also showed that the significant difference was between scenario 1 and 2. It could very well be that social score in scenario 1 was better than in scenarios 2 and 3 because of a novelty effect. For most participants it was a new experience for them when interacting with the robot for the first time. As seen with the demographic variables in Appendix E, 23 out of 30 participants self-reported never having worked with a robot before. It could be that after the first game the novelty effect wore off already, resulting in the sharp decline between levels 1 and 2. And because the novelty has worn off by the second game, the difference between scenario 2 and 3 was not as big.

scenario	time	hints
1	5093	52.7
2	6655	70.26
3	1394	14.8
all	4599	48.67

Table 6.1: Variances of time and hints grouped by scenario

6.1.3. Demographic variables

Because the multivariate analysis with the demographic variables was not done, the results of this analysis can not be discussed. What can be discussed however is what the effect of the demographic variables could have been. The questions along with the distribution of responses can be found in Appendix E. As shown in the Appendix most participants self-reported to have never worked with robots before which could explain the possible novelty effect going on with the social score. Another interesting thing to see is when the performance measures are grouped by completed education as shown in table 6.2. It looks like that while participants with a higher education took less time and needed less hints, this trend did not continue with accuracy. In fact, the highest educated people had the lowest accuracy while the lowest educated had the highest. Actual conclusions shouldn't be drawn from this as the vwo category only contained one person, while the wo master category contained three. Another rather interesting thing to see is the effect pet ownership seemed to have on the mental model and perceptions towards the robot. Pet ownership was divided into three categories: no pets owned, pets owned which are capable of joint attention such as dogs and goats, and other pets. For the total score there was a significant difference whether someone owned a pet or not but it did not matter as much whether the pets were capable of joint attention. This distinction is a bit more visible with social score. It seems as though pet owners had a more positive mental model towards the robot and saw the robot as more of its own entity with its own mind.

education	n	time mean	hints mean	accuracy mean
vwo	3	146.2	15.33	73.33
mbo	12	118.1	14.83	66.67
hbo	66	86.1	10.64	70
wo master	9	64.69	8.333	66.67

Table 6.2: Means of time, hints and accuracy grouped by completed education

pet owner	n	mean total score	mean social score
no	18	-2.611	-4
yes, no JA	42	11.69	-1.19
yes, yes JA	30	11.87	-0.133

Table 6.3: Means of total and social score grouped by pet ownership

6.2. Limitations

While the experiment was performed with adults, the ultimate goal of this project was to develop a system that can be used to train and potentially evaluate joint attention skills in children, especially children with ASD. For this thesis project however, it was both ethically and time-wise not feasible to actually perform the experiments with children. Children are also not as good as adults at putting feelings and opinions into words, which was needed for this thesis project. And because this system is a prototype, a lot of feedback was needed. Adults can better express themselves as to what worked and did not work for them as opposed to children. While using adults for this research project was not a bad choice because of the aforementioned reasons, this did make the tested population very different from the target population: children with ASD. This lowers the external validity of this research.

6.2.1. Limitations of the experiment

One limitation in the experiment was that people might have been reacting to the change in movement of the robot instead of reacting to the hint behaviour. Instead of looking out for the robot to display its hint, they could have been listening to when the movements of the robot changed since the motors of the robot are fairly loud. If this were the case, then this could have diminished or even neutralised the effect of the level of jointness in the hint. This is because for every level the hint would mean a change in movement from the search pattern. The design of the game did try to take this into account by having the robot start searching from a random corner of the game each search pattern, making the search pattern and therefore the sound pattern the robot made different depending on which object the robot was looking for. One way this limitation could be better accounted for would be to replay the game with the same set up but have the robot make use of the idle behaviours this time. Another limitation was that the robot's behaviour was still quite mechanical. There was not enough time to implement a completely natural set of behaviours for the robot. While effort was expended to make the movements of the robot less mechanical and more lifelike, it still felt mechanical to most participants. This may have reduced the effect of joint attention level on the total and social scores used to answer SRQ 2 and 3 but at this point more research is needed to verify that.

6.2.2. Limitations of the system

One limitation of the system was that the eye cameras of the MiRo-E are capable of registering infrared light. The pupil labs glasses use infrared light to illuminate the eyes for their eye cameras, this way the difference in refraction between the pupil and the cornea is better detectable. However, when someone wore the pupil labs glasses, this caused their face to become a bright white dot in the vision of the MiRo-E. This prevented the system from recognizing the face of the participant. This issue was solved quickly by making the robot capable of tracking the brightest spot in its field of view instead of faces. But when the robot tracks the brightest spot instead of a face, the lighting in the room must be controlled so that nothing brighter than the face is in the robot's field of view. This meant that the experiment location was quite dark which for most participants was no issue. A few participants however had a little trouble with seeing in the darkened room. This was not enough to not be able to participate but it may have affected their performance.

6.2.3. Limitations of the analysis

The first limitation in the analysis was that the non-parametric Friedman tests where significant differences were found were under-powered. The effect sizes calculated in chapter 5 were so small that according to G*Power there needed to be at least 106 participants to achieve a power of 80%. This number is a lot larger than the 30 people that did participate. A second limitation is heavily skewed demographic variables. When filling in the answers for the pre or post questionnaire, almost all answers were multiple choice. It can be seen in Appendix E, but almost all the demographic variables which had more than two choices were heavily skewed towards a single choice. For some variables a single choice contained 75% or more of the answers. For example the experience with robots demographic variable had five possible answers, one answer had 76.67% of the responses. A third limitation was that some demographic variables were intentionally not recorded(age, gender) for privacy reasons. But this does cause any results to not be easily generalised to other populations. Since this system was intended as a prototype with the intention that more research would follow, the researcher found the privacy of the participants to be more important than the ease of generalisation to other populations especially since the population used is not the intended population. Finally, as mentioned in section 2.4.3 the answer scale of the ASA questionnaire used in this research project was not entirely the same as how the answer scale was originally developed. While both the original scale and the scale used in this research both use a 7-point Likert scale, the labels are a bit different. In the original scale participants gave an answer of -3 to 3, with -3 meaning disagree, 3 meaning agree and 0 meaning neither agree nor disagree. The scale used in this research makes use of labels for answers, but with the same scoring attached to those labels. See table 2.1 in chapter 2 for how the scoring and labels of the two works relate.

7

Conclusions & Future work

7.1. Conclusions

7.1.1. Sub research questions

In order to answer the main research question, it was divided into three sub research questions which were answered separately. The first sub research question was as follows:

SRQ 1: What is the effect of the level of joint attention on task performance?

In order to measure the task performance, a game was played and the time needed to complete the game, the amount of hints needed, and the accuracy of the participants were recorded. The analysis showed that the hypothesis for this sub research question was supported, so a conclusion could be drawn. The answer to SRQ 1 is: the effect of joint attention on task performance is that joint attention increases task performance.

The second sub research question was:

SRQ 2: What is the effect of the level of joint attention on the human's mental model of the robot?

In order to measure the mental model, the ASA questionnaire described in chapter 2 was used. The participants were asked to fill in the questionnaire after each game with the same questions each time. This questionnaire consisted of multiple subscales. In order to prevent survey fatigue only the most relevant four subscales were used. The used subscales measured if the participants liked the robot, if they thought the robot was intentional, if they thought the robot had a social presence and if they found the robot to be attentive. Together these four subscales measured the mental model the participant had towards the robot. The analysis of whether joint attention level had a significant effect on a human's mental model towards the robot found non-significant results. This research found no evidence to support or reject the hypothesis. So the answer to SRQ 2 is inconclusive.

The third and final sub research question and hypothesis read:

SRQ 3: What is the effect of the level of joint attention on the perception of a robot as its own entity with its own mind?

In order to measure the perception of the robot as its own entity, data from the same questionnaire used for SRQ 2 was used here, albeit only one of the four subscales. The subscale used was the social subscale, where the questions revolved around whether the robot was perceived as a social entity. Just like SRQ 2 the results and analysis whether joint attention level had a significant effect on the social score were non-significant, so the hypothesis for SRQ 3 can't be rejected nor is it supported. The answer to SRQ 3 is therefore inconclusive.

7.1.2. Main research question

Now that the three sub research questions have an answer, albeit said answer is inconclusive for two sub research questions, a conclusion can be drawn for the main research question. The main research question and its hypothesis were defined as:

> **Main RQ: What is the effect of the level of joint attention on the interaction with a human?**
> **Main hypothesis: An increase in the robot's joint attention level leads to a more positive interaction between a robot and a human.**

Interaction with a human was divided into three parts: human's task performance, the human's mental model of the robot and the human's perception of the robot as its own entity. For two of these parts: the human's mental model of the robot and the human's perception of the robot as its own entity it was concluded that it can neither be supported or rejected that joint attention had an significant effect. For the human's task performance however, it was concluded the hypothesis was supported. The main research question was divided into three sub research questions, one sub research question had its hypothesis supported while the other two were inconclusive. This means that the main hypothesis is partially supported.

With the main hypothesis partially supported, an answer to the main research question can be formed. The objective measures time and hint showed improvement as the level of joint attention went up, i.e. less time and hints were needed to complete a game. And although it suggested the same for accuracy when looking at just the data, the statistical analysis showed that the difference in means between the joint attention levels was not significant. As mentioned in chapter 5, the hypothesis for SRQ 1 was supported. This is quite different from the subjective measures total and social score. The data itself showed that these measures remained stable between joint attention levels, already suggesting that joint attention level did not have an impact on these subjective measures. The analysis did not disprove this, for both total and social score no significant difference in the mean was found and as mentioned in chapter 5, the hypotheses for SRQ 2 and 3 were neither supported nor rejected. This difference between objective and subjective measures may be because higher levels of joint attention involved looking at the object more often, and thus gave more information to the participants allowing them to form a guess in less time and with less hints. For the mental model and perception of the participant towards the robot however, higher levels of joint attention did not make enough of a difference. Humans form a mental model before even interacting with the robot [42, 43], it could be that the behaviour of the robot was not life-like enough to influence the mental models and perceptions regardless of joint attention level.

So while the hypotheses for SRQ 2 and 3 were not rejected but also not supported, the hypothesis for SRQ 1 was supported. This means that this research found an answer to the main research question: the level of joint attention has at least one effect on the interaction with a human: it improves task performance.

7.2. Experiment variation

7.2.1. Game variation

Another way the game could be played is that instead of having the robot move from object to object, have it start in a neutral position and then use one of the three levels of joint attention to point to a single object. The participant must then guess which object the robot was looking for. This game could also be used to test if participants know which level of joint attention is used by making the participant not only guess the object, but also the joint attention level. By playing the game this way the issue where participants are potentially only reacting to the sounds and noises of the robot could be combated. This is because the change in noise does not indicate a hint to be exploited, but rather the start of the game.

7.2.2. First scenario only

The main analysis focused on the levels of joint attention the robot could use and the effect this had had on a human's task performance, their mental model and their perception of the robot. This analysis revealed that while joint attention level had an effect on the performance, no significant effect on the human's mental model and perception of the robot was found. The exploratory analysis explored what effect scenario itself had on the outcome variables. This analysis showed that scenario had an significant effect on hints, accuracy and social score, but not on time and total score. When looking at the means of the data by joint attention level and grouped by scenario, something interesting becomes visible. When looking only at the first scenario in figures 7.1 through 7.5 below, the trends visible are mostly what was expected at first, but these trends break from the second scenario onward. The analysis already revealed that a learning effect might have affected the experiment. This is why it might be interesting to repeat this experiment, but then have participants only play one game. Like a between-subjects version of this experiment. Every participant only plays one game with a random level of joint attention and fills in the questionnaire. Because the questionnaire does not need to be repeated and therefore questionnaire fatigue is less of an issue, a larger part of the ASA questionnaire could be used.

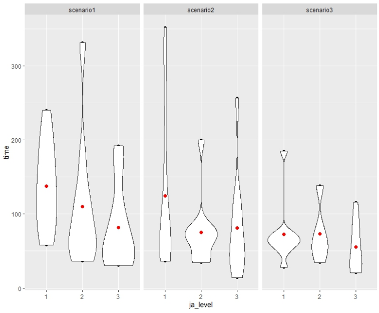


Figure 7.1: Mean time by joint attention level grouped by scenario

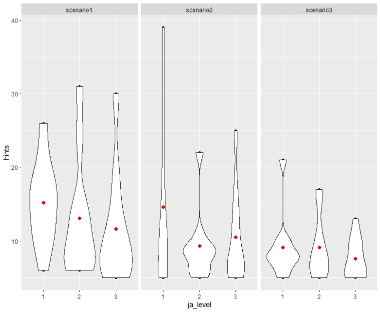


Figure 7.2: Mean hints by joint attention level grouped by scenario

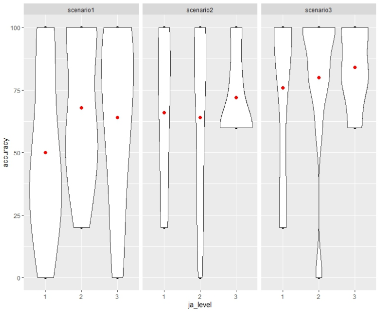


Figure 7.3: Mean accuracy by joint attention level grouped by scenario

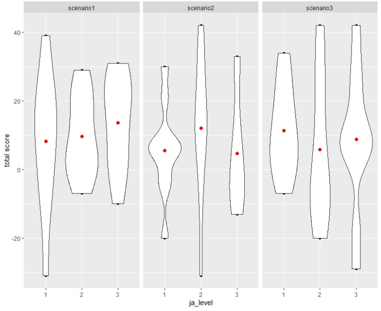


Figure 7.4: Mean total score by joint attention level grouped by scenario

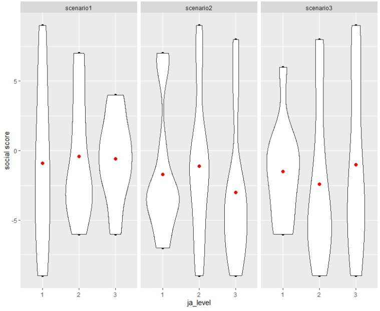


Figure 7.5: Mean social score by joint attention level grouped by scenario

7.3. Future research

This system was very much a prototype, made to show there is promise here but there is a lot of room for improvement. Future work could address the limitations listed in chapter 6, fixing issues such as making the movements less mechanical and having more participants. Especially the mechanical movements should be made more natural and then see if joint attention level does have an impact on the human's mental model and perception. This is because behaviour is important when it comes to seeing if a robot is perceived as having a mind. In a paper by Wiese et al. [38] it is stated that an agent needs to be viewed as having mind via their behaviour before appearance even matters. That the robot is perceived as having a mind has an impact on the degree to which gaze is followed on an agent [13].

Future research could also replicate this experiment, but then use children as participants as they were the intended focus group. A study by Warren et al. [31] used a robot to help children diagnosed with ASD develop joint attention skills, however the robot they used was a humanoid robot. The replication of this experiment could focus on if this system can help teach children, especially children with ASD, joint attention skills by playing games with them. Furthermore it could also be investigated if children like playing these games. This system could provide a non-intrusive way to teach children these important skills, helping them to better connect socially with their peers. Follow up research could be to develop and research what kind of games work better to teach children joint attention skills. Or investigate if children have a preference for certain games. In a study by van Otterdijk et al. [46] they concluded that the personalisation of the games played with the robot help sustain attention and engagement towards the robot.

Another direction future research with children with ASD could take is developing a non-intrusive method of diagnosing and evaluating ASD in children. Gaze trajectories can differ between children with ASD and those without, and children with ASD sometimes show a spontaneous increase in gaze behaviour in response to robots as opposed to other humans. Because of these reasons, a system could be developed that evaluates the response children have towards the robot and could help indicate whether they have ASD [14].

Future research could also replicate this experiment but then play the same games using different kinds of robots. The system was designed to make it easy to repeat experiments with different robots, this way the impact of the design and appearance could be evaluated. A robot's appearance influences a human's mental model and expectations of the robot [9]. This system could provide an easily replicable way to repeat experiments with different robots to test appearance and design within the context of joint attention. Other contexts could also be possible should this system be expanded upon.

A

Phases & Levels

A.1. Starting Level of common attention

Take two people, Anne and Bart, and a cube as an object on which joint attention will be performed. Anne and Bart are in the same room, and can both see the ball but at this point in time they do not know if the other is also looking at the cube. This means:

- Anne is aware of the cube.
- Bart is aware of the cube.
- Anne is not aware that Bart is aware of the cube.
- Bart is not aware that Anne is aware of the cube.

Thus at the start Anne and Bart have common attention on the cube as defined in chapter 2.

A.2. Common to monitoring attention

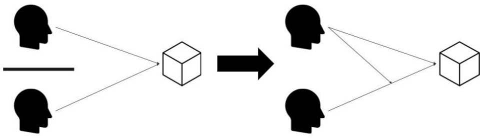


Figure A.1: From common to monitoring attention

Now Anne establishes mutual gaze with Bart. This means that after mutual gaze is established:

- Anne is aware of the cube.
- Bart is aware of the cube.
- Anne is now aware that Bart is aware of the cube.
- Bart is still not aware that Anne is aware of the cube.

This means that Anne and Bart now have monitoring attention on the cube as defined in chapter 2. Bart is not aware that Anne is aware of the cube because he was looking at the cube and not Anne, therefore he did not see Anne looking at the cube. After the establishing of mutual gaze the joint attention level went up from common to monitoring.

A.3. Monitoring to mutual attention

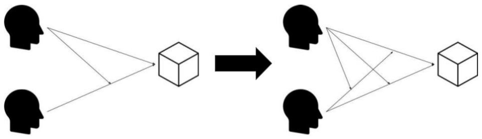


Figure A.2: From monitoring to mutual attention

Now Anne uses referential gaze toward the cube. This means that after the referential gaze is successful:

- Anne is aware of the cube.
- Bart is aware of the cube.
- Anne is aware that Bart is aware of the cube.
- Bart is now aware that Anne is aware of the cube.
- Anne is not certain that Bart is aware that Anne is aware of the cube.
- Bart is not certain that Anne is aware that Bart is aware of the cube.

Anne is not certain that Bart is aware that Anne is aware of the cube because Anne does not know at that point that the referential gaze was successful. This is because Anne is looking at the cube and not at Bart to confirm the success of the referential gaze. And while Anne is aware that Bart is aware of the cube, Bart is not certain of that information because it is possible that Anne did not see the cube when mutual gaze was established. This would mean that because at this point in time Anne is looking at the cube, she does not know Bart is aware of the cube. So Bart is not certain that Anne is aware that Bart is aware of the cube. In reality however Anne is aware that Bart is aware of the cube because she saw him looking at it before mutual gaze was established so Anne and Bart now have mutual attention on the cube as defined in chapter 2. But this uncertainty prevents it from being shared attention as defined in chapter 2.

A.4. Mutual to shared attention

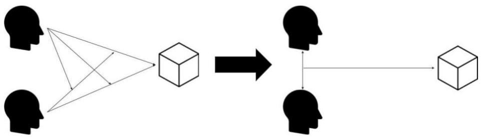


Figure A.3: From mutual to shared attention

Finally, Anne established mutual gaze with Bart again, confirming the uncertainty from the previous step. Because now Anne knows the referential gaze was successful, and Bart now knows Anne saw him looking at the cube. So at this point:

- Anne is aware of the cube.
- Bart is aware of the cube.
- Anne is aware that Bart is aware of the cube.
- Bart is aware that Anne is aware of the cube.
- Anne is now certain that Bart is aware that Anne is aware of the cube.
- Bart is now certain that Anne is aware that Bart is aware of the cube.

Anne and Bart now have shared attention on the cube.

B

Initial data analysis

In this initial data analysis the assumptions for performing ANOVA were checked, it was checked if the residuals were normal distributed, if there was homogeneity of variance and if there was sphericity. Sphericity needed to be checked because the data is from a repeated measures designed experiment. For all measures the method to fit the model in R was: aov(measure ~ ja_level + (1 | participant), data=data). Aov was used because this is an ANOVA wrapper for linear models.

B.1. SRQ1; the performance measures

B.1.1. Time

Aside from plotting the residuals as shown in figures B.1 and B.2 below, a Shapiro-Wilk test was also performed. The test showed that the distribution of the residuals of time departed significantly from normality: W = 0.80904, p = 1.938e-09.

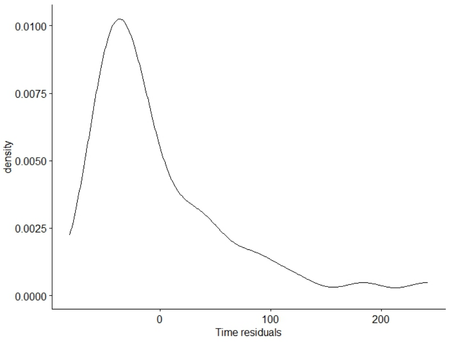


Figure B.1: Density plot of residuals of time

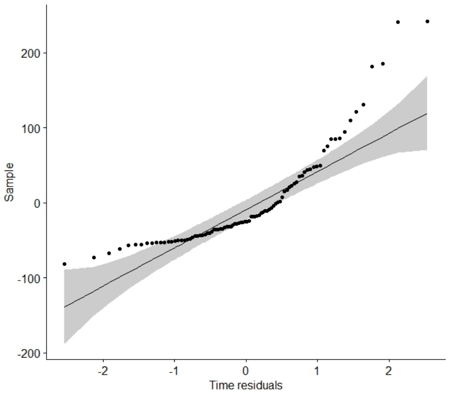


Figure B.2: QQ plot of residuals of time

Next was to check the homogeneity of variance using Levene's Test. Levene's Test showed that there was no significant difference in variance with F(2,87) = 1.0094, p = 0.3687. A visualisation of the variance can be seen in figure B.3 below.

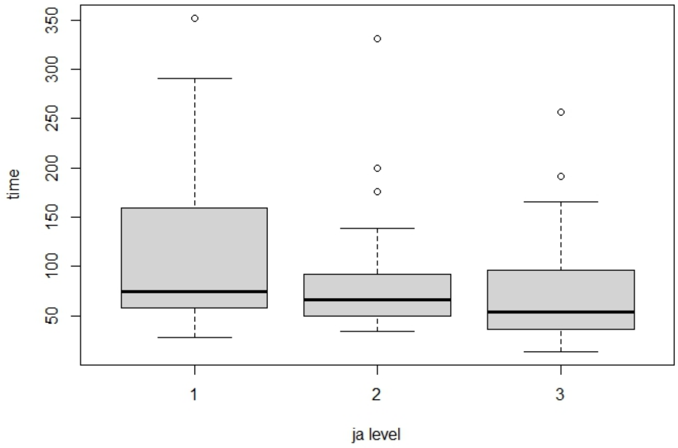


Figure B.3: variances of time

The next step would be to check sphericity but this did not need to be checked since on top of the evidence for violation of the normality assumption, time has multiple outliers as seen in figure 5.7. So Friedman was used instead of ANOVA, with Nemenyi as Post-hoc analysis.

Hints:

Like with the time measure, a Shapiro-Wilk test was performed aside from plotting the residuals as shown in figures B.4 and B.5 below. The test showed that the distribution of the residuals of hints also departed significantly from normality: W = 0.80786, p = 1.785e-09.

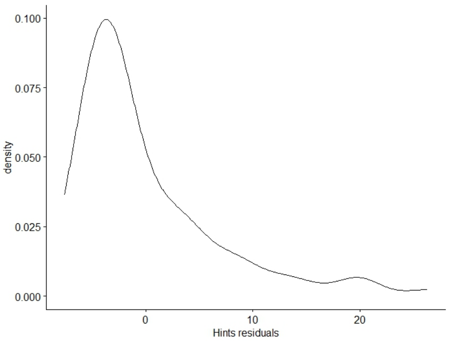


Figure B.4: Density plot of residuals of hints

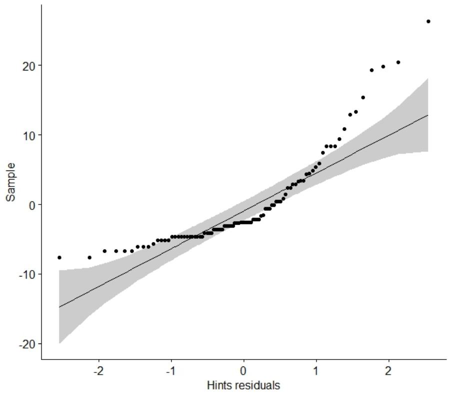


Figure B.5: QQ plot of residuals of hints

The next assumption to check was the homogeneity of variance using Levene's Test. Levene's Test showed that there was no significant difference in variance with F(2,87) = 0.9919, p = 0.375. A visualisation of the variance can be seen in figure B.6 below.

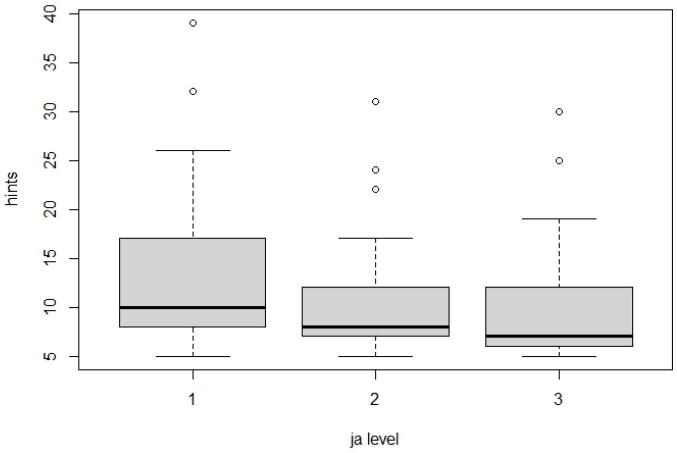


Figure B.6: variances of hints

Just like with time, sphericity did not need to be checked here since hints has outliers as seen in figure 5.8 on top of the evidence that the normality assumption was violated. So like with the time measure, Friedman was used with Nemenyi as Post-hoc analysis.

Accuracy:
As with the measures above, the residuals were plotted as shown in figures B.4 and B.5 below and a Shapiro-Wilk test was also performed. The test showed that the distribution for the residuals of accuracy were not normal distributed: W = 0.91181, p = 1.423e-05.

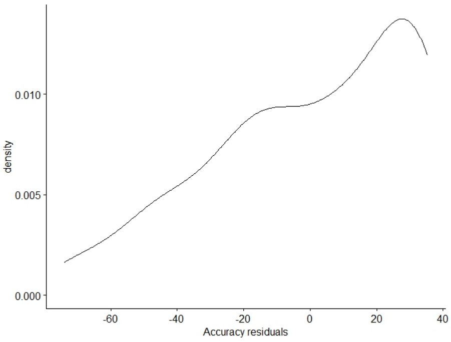


Figure B.7: Density plot of residuals of accuracy

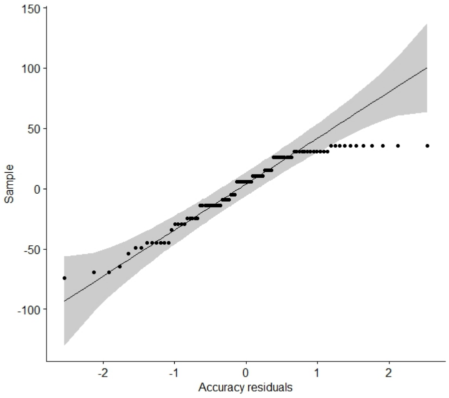


Figure B.8: QQ plot of residuals of accuracy

Like the previous measures the homogeneity of variance was checked using Levene's Test. Levene's Test showed that there was a significant difference in variance with F(2,87) = 3.2444, p = 0.04376. A visualisation of the variance can be seen in figure B.9 below.

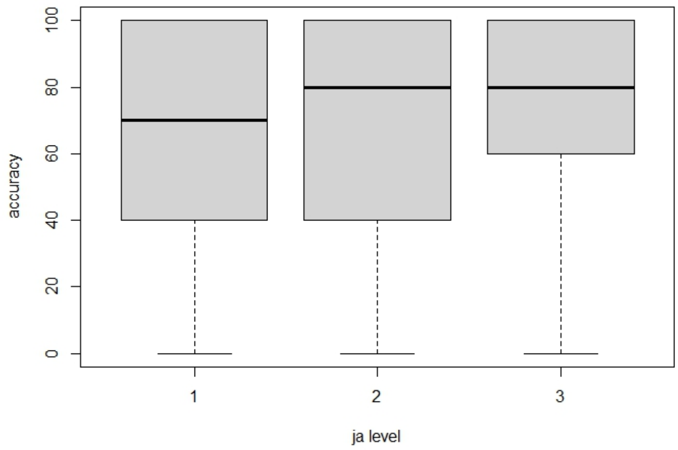


Figure B.9: variances of accuracy

This means that there was evidence that the data for accuracy violated two assumptions, so ANOVA was not used here. Instead Friedman was used here with Nemenyi as Post-hoc analysis.

B.2. SRQ2; the total score

First an initial data analysis was performed just like with the analysis for the first sub research question. The residuals of the linear mixed model were plotted in a density plot and a QQ plot and a Shapiro-Wilk test was performed. The residuals of the data for the total score did not significantly depart from a normal distribution. This is visible in both the density plot and the QQ plot as shown in figures B.10 and B.11 below.

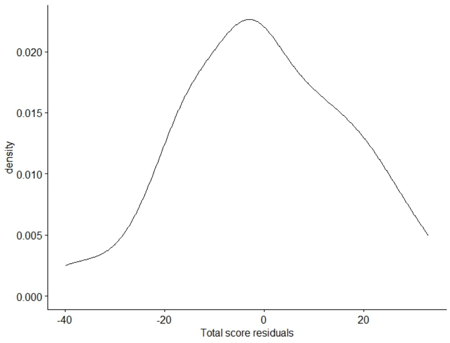


Figure B.10: Density plot of residuals of total score

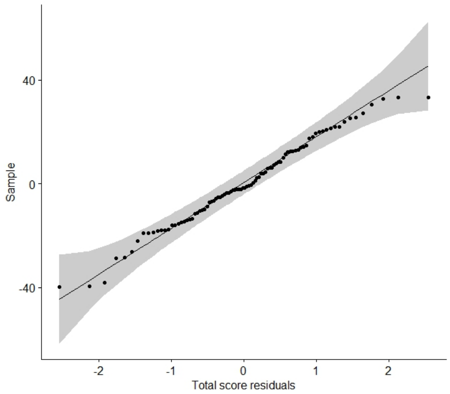


Figure B.11: QQ plot of residuals of total score

A Shapiro-Wilk test was also performed, and this test showed that the difference with a normal distribution is non-significant: W = 0.98558, p = 0.4239. It could then be concluded that there was no indication that the normality assumption was violated. The next assumption to test was the homogeneity of variance using Levene's test. Levene's test showed that the variances did not differ significantly, with F(2,87) = 0.793, p = 0.4557. A visualisation of the variances are shown in figure B.12 below.

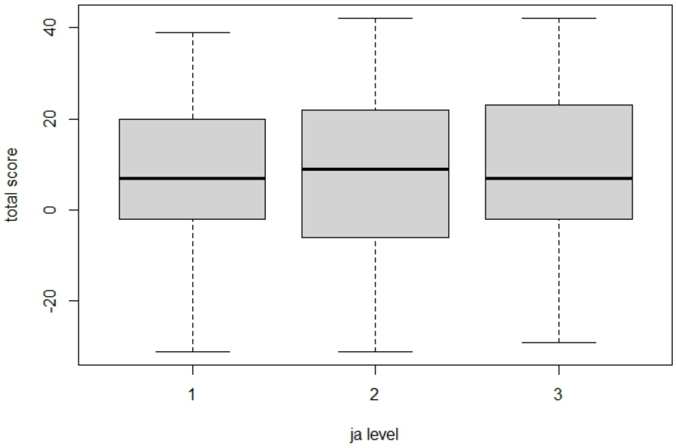


Figure B.12: variances of total score

Sphericity was tested using Mauchly's Test of Sphericity, and this test showed that the data has sphericity with F(2,87) = 0.13, p = 0.879. Since no evidence was found for any violation of the necessary assumptions and there were no outliers as shown in figure 5.10, repeated measures ANOVA was used.

B.3. SRQ3; the social score

First an initial data analysis was performed just like with the first sub research question. The residuals of the linear mixed model were plotted in a density plot and a QQ plot and a Shapiro-Wilk test was performed. The residuals of the data for the total score were not normally distributed. This is visible in both the density plot and the QQ plot as shown in figures B.13 and B.14 below.

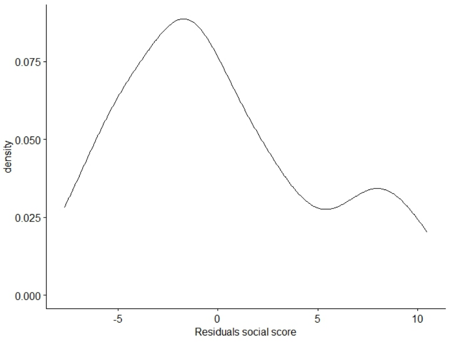


Figure B.13: Density plot of residuals of social score

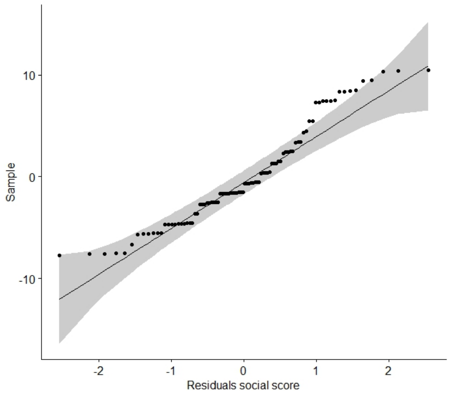


Figure B.14: QQ plot of residuals of social score

Furthermore the Shapiro-Wilk test also showed the distribution has a significant departure from normality: W = 0.94159, p = 0.0005251. So it was safe to say that there was evidence to support that the normality assumption was violated. But just like with the first sub research question, the other assumptions were still tested since ANOVA is quite robust when it comes to normality violations. The next assumption to test was the homogeneity of variance using Levene's test. Levene's test showed that there was no significant difference in variances, with F(2,87) = 0.1262, p = 0.8816. The figure B.15 below shows a visualisation of the variances.

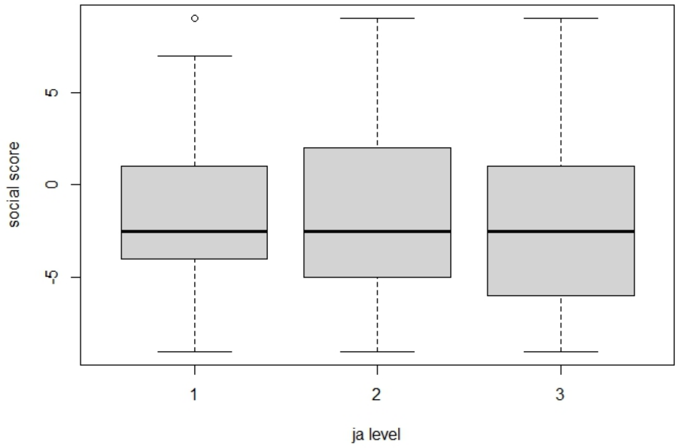


Figure B.15: variances of social score

There was no indication that the assumptions for ANOVA were violated, but like SRQ 2, Sphericity needed to be tested before repeated measures ANOVA could be used. Mauchly's Test of Sphericity showed that the data has sphericity with F(2,87) = 0.13, p = 0.879

All the assumptions necessary were tested and there was only evidence that the normality assumption was violated. Furthermore social score has no outliers as seen in figure 5.11, so repeated measures ANOVA was used.

C
Exploratory analysis

C.1. Scenario as IV

Before any analysis could be performed, an initial data analysis needed to be performed. First the residuals for the performance measures time, amount of hints and accuracy were checked to see if the data was normally distributed. Two linear mixed models were fitted on this data and the residuals were plotted in both a density and a QQ plot. And just to be sure a Shapiro-Wilk test was also performed on this data. For all measures the method to fit the model in R was:
aov(measure ~ scenario + (1 | participant), data=data)
The next step was to check the homogeneity of variance using Levene's test. And finally sphericity was checked using Mauchly's test of sphericity. Just like the main analysis, ANOVA was used in the analysis if there was only indication for the violation of the normality assumption and there were no outliers. If there was indication for the violation of two or more assumptions, or there was indication that the normality assumption is violated together with the presence of outliers, ANOVA was not used. In this case the non-parametric Friedman test was used instead.

C.1.1. The performance measures

Time

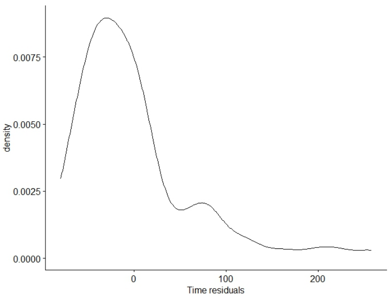


Figure C.1: Density plot of residuals of time

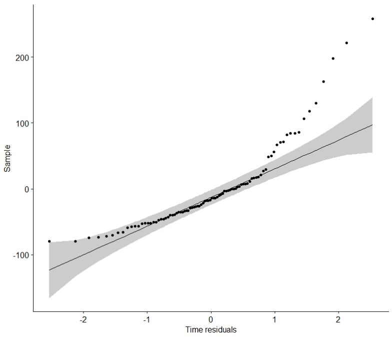


Figure C.2: QQ plot of residuals of time

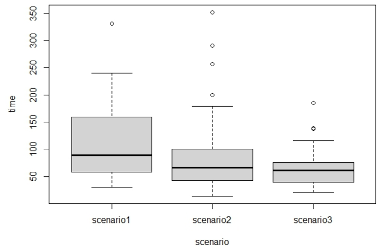


Figure C.3: Variances of time

Figures C.1 and C.2 above show the residuals visualised in both a density and QQ plot. The Shapiro-Wilk test showed that the distribution of the residuals of time departed significantly from normality: W = 0.83893, p = 1.754e-08. Next was to check the homogeneity of variance using Levene's Test. Levene's Test showed that there was no significant difference in variance with F(2,87) = 2.5728, p = 0.08212. A visualisation of the variances can be found in figure C.3. The last assumption to be tested before repeated measures ANOVA could be used was sphericity. Sphericity was tested using Mauchly's Test of Sphericity, and it showed that there was indication that the sphericity assumption was violated with F(2,58) = 5.888, p = 0.005. So with indication that 2 assumptions were violated ANOVA was not used and Friedman was used instead.

Friedman showed that for time there was no significant difference between scenarios with X2(2) = 2.6, p = 0.2752.

Hints:

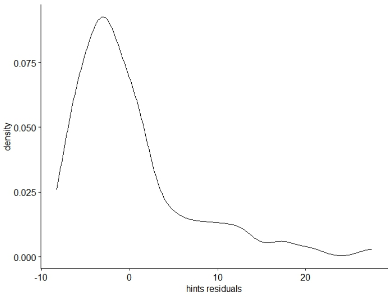


Figure C.4: Density plot of residuals of hints

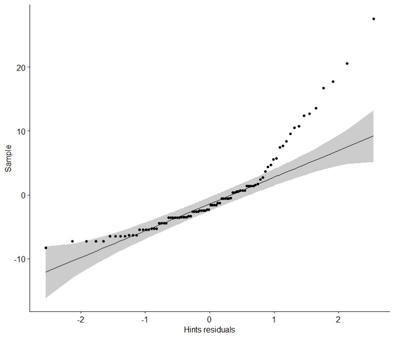


Figure C.5: QQ plot of residuals of hints

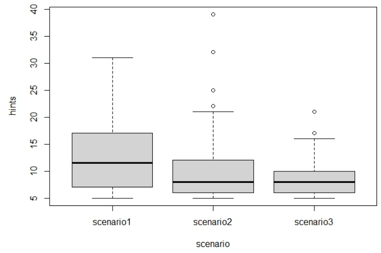


Figure C.6: Variances of hints

Figures C.4 and C.5 above show the residuals visualised in both a density and QQ plot. The Shapiro-Wilk test showed that the distribution of the residuals of time departed significantly from normality: W = 0.83793, p = 1.623e-08. Next was to check the homogeneity of variance using Levene's Test. Levene's Test showed that there was no significant difference in variance with F(2,87) = 2.5709, p = 0.08226. A visualisation of the variances can be found in figure C.6 The last assumption to be tested before repeated measures ANOVA could be used was sphericity. Sphericity was tested using Mauchly's Test of Sphericity, and it showed that there was indication that the sphericity assumption was violated with F(2,58) = 7.608, p = 0.001. So with indication that 2 assumptions were violated ANOVA was not used and Friedman was used instead.

Friedman showed that for hints there was a significant difference between scenarios with X2(2) = 8.7664, p = 0.01249. Post-hoc analysis using the Nemenyi test between scenarios revealed statistically significant differences between scenario 1 and 3(q = 3.925, p = 0.015219).

Accuracy:

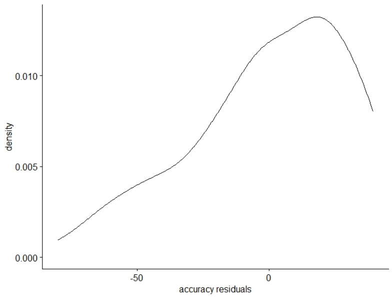


Figure C.7: Density plot of residuals of accuracy

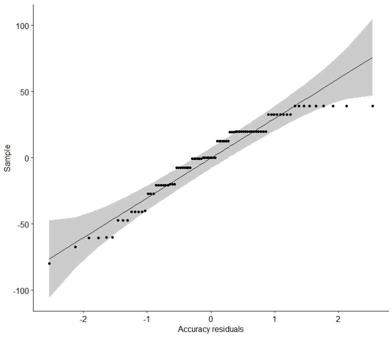


Figure C.8: QQ plot of residuals of accuracy

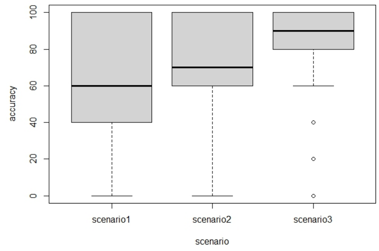


Figure C.9: Variances of accuracy

Figures C.7 and C.8 above show the residuals visualised in both a density and QQ plot. The Shapiro-Wilk test showed that the distribution of the residuals of time departed significantly from normality: W = 0.93318, p = 0.0001761. Next was to check the homogeneity of variance using Levene's Test. Levene's Test showed that there was no significant difference in variance with F(2,87) = 1.1708, p = 0.315. A visualisation of the variances can be found in figure C.9 The last assumption to be tested before repeated measures ANOVA could be used was sphericity. Sphericity was tested using Mauchly's Test of Sphericity, and it showed that there was indication that the sphericity assumption was violated with F(2,58) = 6.603, p = 0.003. So with indication that 2 assumptions were violated ANOVA was not used but Friedman was used instead.

Friedman showed that for time there was no significant difference between scenarios with X2(2) = 11.062, p = 0.003963. Post-hoc analysis using the Nemenyi test between scenarios revealed statistically significant differences between scenario 1 and 3(q = 3.651, p = 0.026564).

C.1.2. The total score

First an initial data analysis is performed just like with the performance measures.

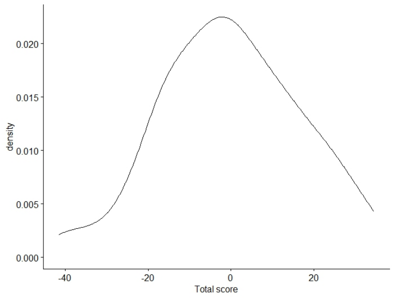


Figure C.10: Density plot of residuals of total score

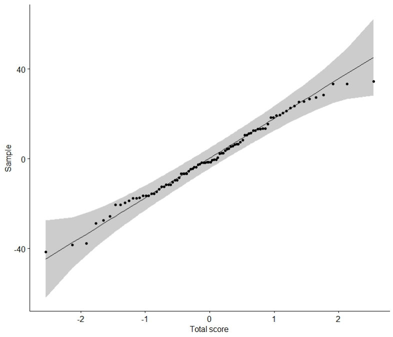


Figure C.11: QQ plot of residuals of total score

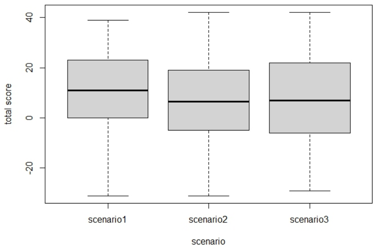


Figure C.12: Variances of total score

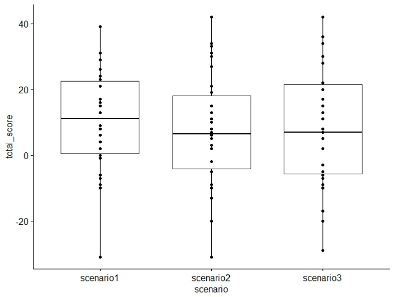


Figure C.13: Outliers of total score

Like the analysis of the performance measures, a density and QQ plot were plotted with the residuals of total score and a Shapiro-Wilk test was performed. The plots can be found in figures C.10 and C.11. The test showed that the difference with a normal distribution was non-significant: W = 0.9886, p = 0.6282. It could be concluded that there was no indication that the normality assumption was violated, so the next assumption to test was the homogeneity of variance using Levene's test. Levene's test showed that the variances did not differ significantly, with F(2,87) = 0.3221, p = 0.7255. A visualisation of the variances can be found in figure C.12 Sphericity was tested using Mauchly's Test of Sphericity, and this test showed that the data had sphericity with F(1.4,40.55) = 2.03, p = 0.156. Since no evidence was found for any violation of the necessary assumptions, and there were no outliers as shown in figure C.13, repeated measures ANOVA was used.

ANOVA showed that for the total score there were no significant differences in the mean between scenarios with F(2,87) = 0.2502, p = 0.7792.

C.1.3. The social score

Just like the analyses above, an initial data analysis was performed first.

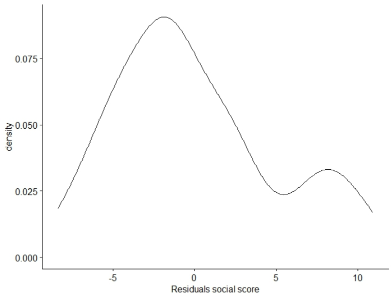


Figure C.14: Density plot of residuals of social score

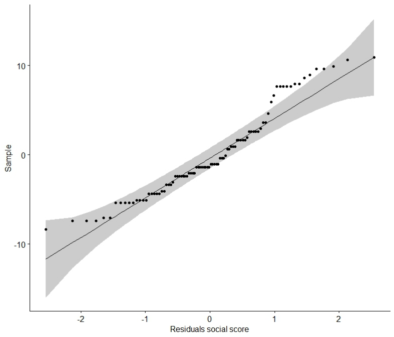


Figure C.15: QQ plot of residuals of social score

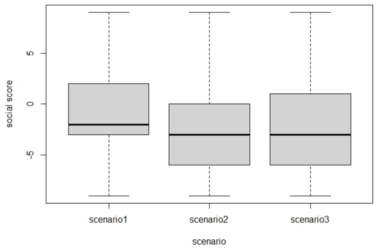


Figure C.16: Variances of social score

A density and QQ plot were plotted with the residuals of social score and a Shapiro-Wilk test was performed. The plots can be found in figures C.14 and C.15. The test showed that the distribution had a significant departure from normality: W = 0.94133, p = 0.000507. It could be concluded that there was an indication that the normality assumption was violated. The next assumption to test was the homogeneity of variance using Levene's test. Levene's test showed that the variances did not differ significantly, with F(2,87) = 0.1511, p = 0.86. A visualisation of the variances can be found in figure C.16 Sphericity was tested using Mauchly's Test of Sphericity, and this test showed that there was an indication the data had no sphericity with F(1.4,38.79) = 5.89, p = 0.013. Since evidence was found that two of the assumptions necessary were violated, repeated measures ANOVA was not used and Friedman was used instead.

This test showed that for social score there were significant differences in the mean between scenarios: X2(2) = 9.9277, p = 0.006986. Post-hoc analysis using Nemenyi showed that there was a statistically significant difference between scenario 1 and 2(q = 3.651, p = 0.026564).

D

Data tables & R outputs

D.1. Data exploration tables

D.1.1. Time

ja_level	n	mean	sd
1	30	111.6	79.88
2	30	86.21	62.3
3	30	72.81	55.2

Table D.1: Time grouped by ja level

scenario	n	mean	sd
1	30	109.9	71.36
2	30	93.69	81.58
3	30	67.03	37.34

Table D.2: Time grouped by scenario

D.1.2. Hints

ja_level	n	mean	sd
1	30	12.97	8.327
2	30	10.5	6.146
3	30	9.9	6.065

Table D.3: Hints grouped by ja level

scenario	n	mean	sd
1	30	13.3	7.259
2	30	11.47	8.382
3	30	8.6	3.847

Table D.4: Hints grouped by scenario

D.1.3. Accuracy

ja_level	n	mean	sd
1	30	64	32.2
2	30	70.67	32.26
3	30	73.33	24.82

Table D.5: Accuracy grouped by ja level

scenario	n	mean	sd
1	30	60.67	33
2	30	67.33	28.52
3	30	80	27.79

Table D.6: Accuracy grouped by scenario

D.1.4. Total score

ja_level	n	mean	sd
1	30	8.433	15.54
2	30	9.2	18.24
3	30	9.033	17.01

Table D.7: Total score grouped by ja level

scenario	n	mean	sd
1	30	10.53	15.22
2	30	7.467	17.09
3	30	8.667	18.31

Table D.8: Total score grouped by scenario

D.1.5. Social score

ja_level	n	mean	sd
1	30	-1.367	4.745
2	30	-1.3	5.114
3	30	-1.533	4.995

Table D.9: Social score grouped by ja level

scenario	n	mean	sd
1	30	-0.633	4.522
2	30	-1.933	5.058
3	30	-1.633	5.163

Table D.10: Social score grouped by scenario

D.2. Rstudio outputs

Listing D.1: Outliers

```
time:
----------------------------------------------------------------------
 participant   scenario    ja_level   time    is.outlier   is.extreme
------------- ----------- ---------- ------- ------------ ------------
      8        scenario1      1       240.2      TRUE         FALSE

     16        scenario2      1       291.3      TRUE         FALSE

     15        scenario2      3       256.6      TRUE         FALSE

     24        scenario1      2       331.5      TRUE          TRUE

     24        scenario2      1       351.6      TRUE          TRUE
----------------------------------------------------------------------
hints:
----------------------------------------------------------------------
 participant   scenario    ja_level   hints   is.outlier   is.extreme
------------- ----------- ---------- ------- ------------ ------------
     10        scenario1      2        24        TRUE         FALSE

      8        scenario1      1        26        TRUE         FALSE

     16        scenario1      3        30        TRUE         FALSE

     16        scenario2      1        32        TRUE         FALSE

     15        scenario2      3        25        TRUE         FALSE

     24        scenario1      2        31        TRUE         FALSE

     24        scenario2      1        39        TRUE          TRUE
----------------------------------------------------------------------
```

Listing D.2: Correlations

```
-------------------------------------------------------------------------
    var1            var2          cor      statistic      p       method
-------------- -------------- --------- ----------- ---------- ---------

     time           hints          0.9        12.19     3.46e-34   Kendall

     time         accuracy        -0.33      -4.176     2.96e-05   Kendall

     time        total_score     -0.0045    -0.06277      0.95     Kendall

     time        social_score     0.052      0.7068       0.48     Kendall

     time          ja_level       -0.2       -2.481      0.0131    Kendall

     time          scenario       -0.2       -2.407      0.0161    Kendall

    hints          accuracy       -0.37      -4.544     5.51e-06   Kendall

    hints        total_score      0.001      0.01402     0.989     Kendall

    hints        social_score     0.072      0.9424      0.346     Kendall

    hints          ja_level       -0.16      -1.879      0.0602    Kendall

    hints          scenario       -0.23      -2.717      0.0066    Kendall

   accuracy      total_score      -0.13      -1.633      0.102     Kendall

   accuracy      social_score     -0.15      -1.808      0.0706    Kendall

   accuracy        ja_level       0.068      0.7536      0.451     Kendall

   accuracy        scenario       0.23        2.555      0.0106    Kendall

 total_score     social_score     0.68        9.084     1.05e-19   Kendall

 total_score       ja_level       0.004      0.04817     0.962     Kendall

 total_score       scenario      -0.043      -0.515      0.607     Kendall

 social_score      ja_level      -0.018      -0.2156     0.829     Kendall

 social_score      scenario      -0.097      -1.149      0.251     Kendall
```

E

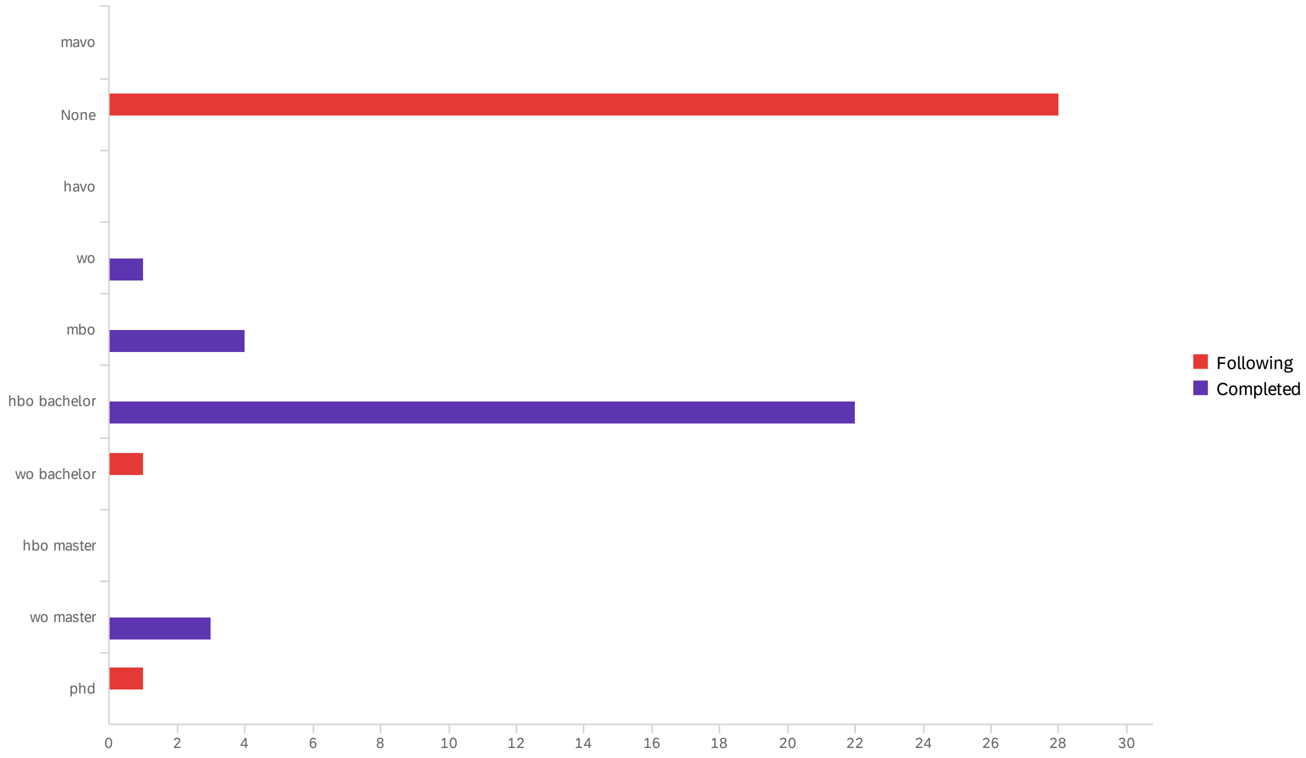
demographic variables

demographic variables

surveys

October 11, 2023 5:17 PM CEST

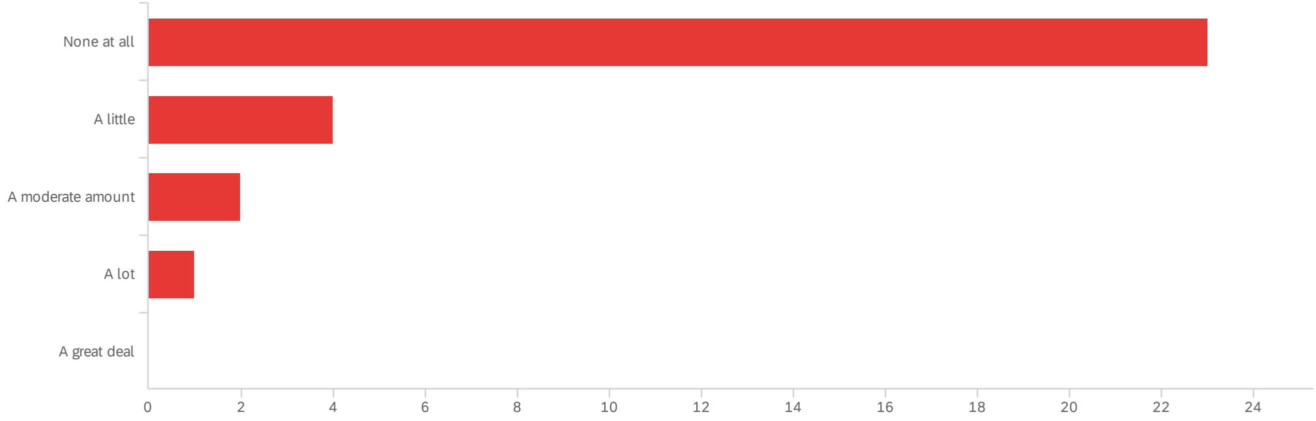
Pre-Q2 - What is the highest education you are following/ have completed?



#	Field	None		wo		mbo		hbo bachelor		wo bachelor		wo master		phd	
1	Following	93.33%	28	0.00%	0	0.00%	0	0.00%	0	3.33%	1	0.00%	0	3.33%	1
2	Completed	0.00%	0	3.33%	1	13.33%	4	73.33%	22	0.00%	0	10.00%	3	0.00%	0

Showing rows 1 - 2 of 2

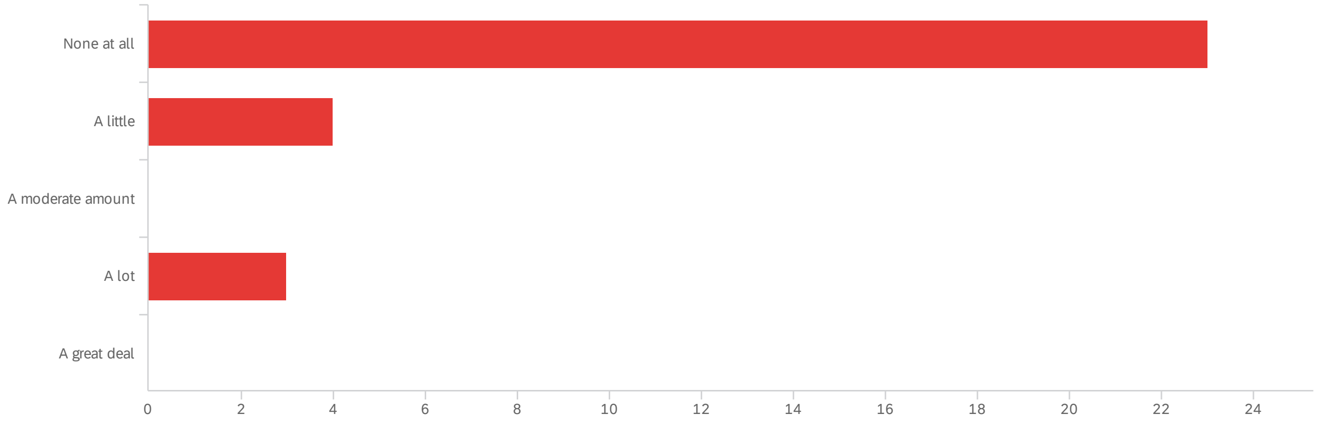
Pre-Q3 - How much experience do you have working with robots?



#	Field		Choice Count
1	None at all	76.67%	23
2	A little	13.33%	4
3	A moderate amount	6.67%	2
4	A lot	3.33%	1
5	A great deal	0.00%	0
			30

Showing rows 1 - 6 of 6

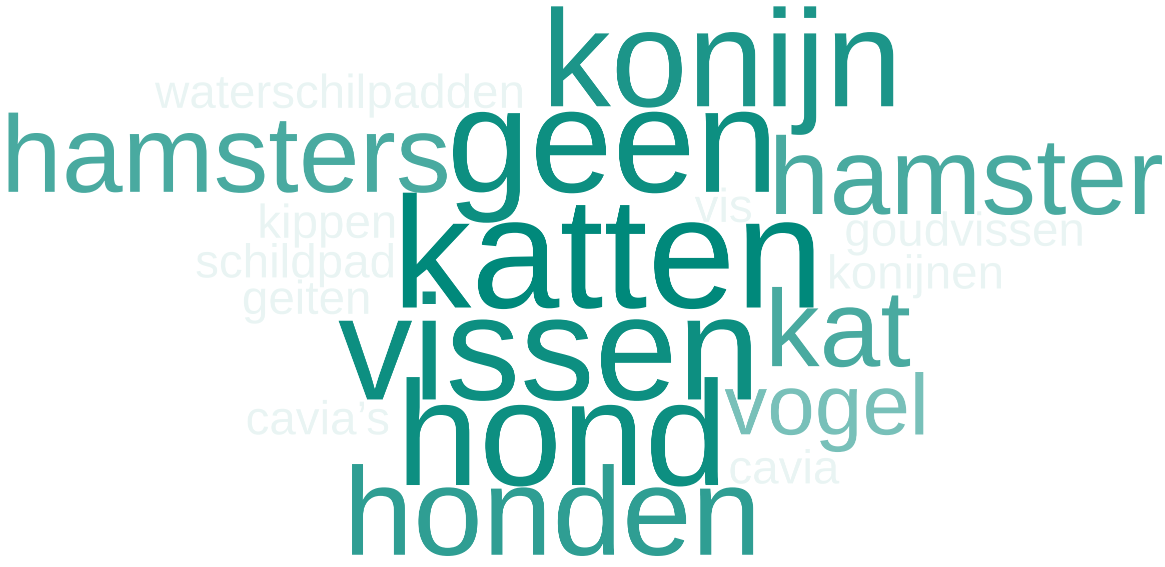
Pre-Q4 - How much experience do you have working with animals?



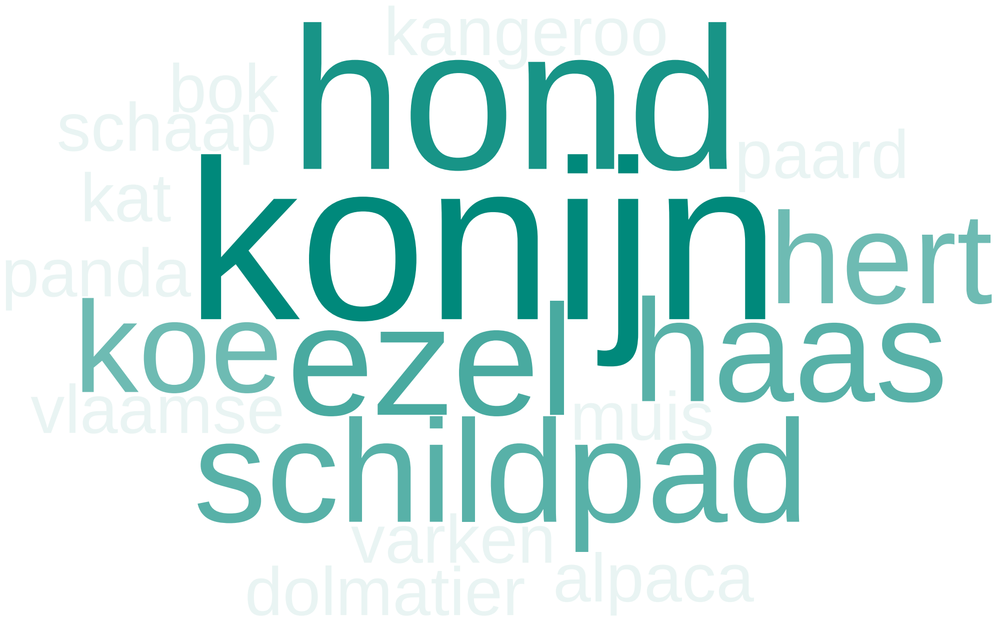
#	Field	Choice Count	
1	None at all	76.67%	23
2	A little	13.33%	4
3	A moderate amount	0.00%	0
4	A lot	10.00%	3
5	A great deal	0.00%	0
			30

Showing rows 1 - 6 of 6

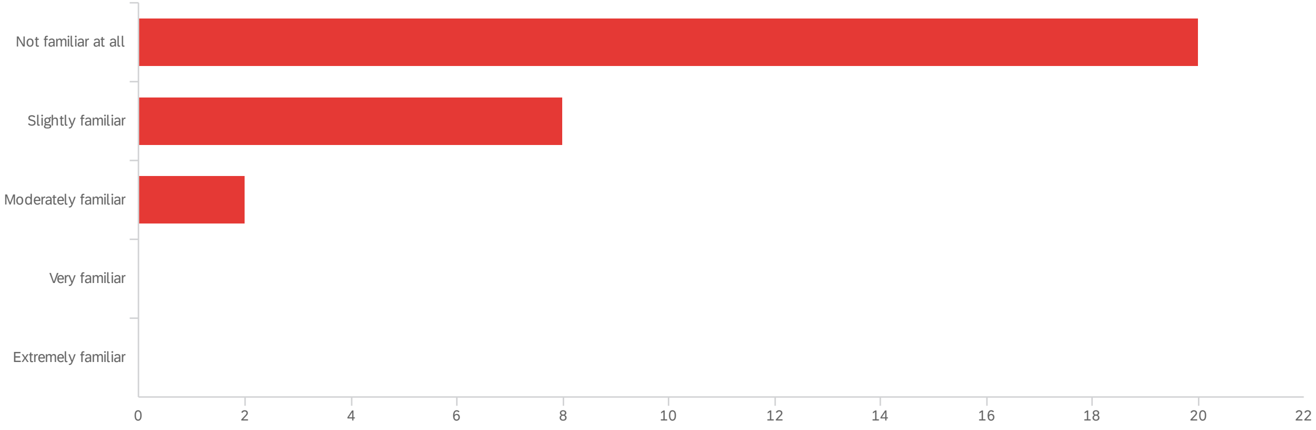
Pre-Q5 - What pets have you owned? List per line the kind of pet and the amount you have had, or None if you never had pets.



Pre-Q6 - What kind of animal do you think the Miro resembles? Multiple animals are allowed



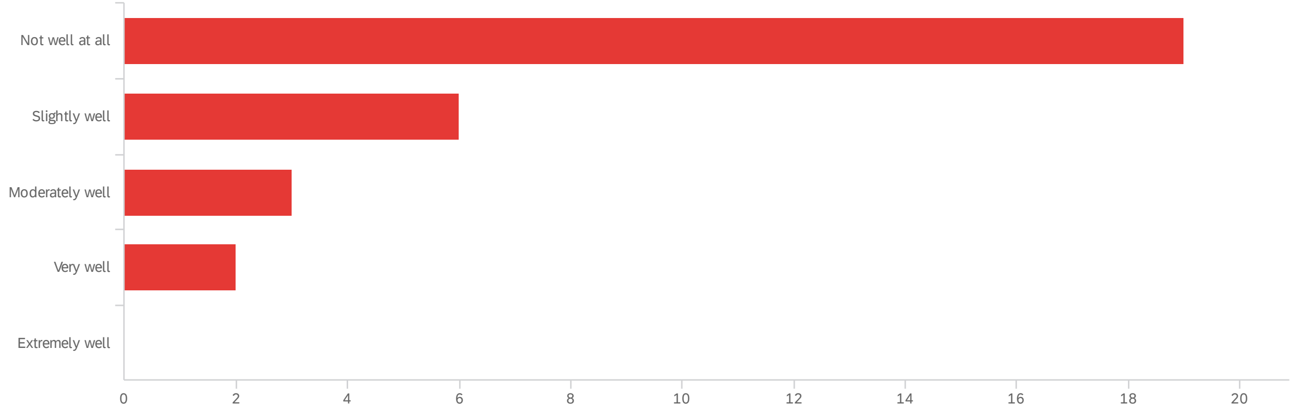
Post-Q1 - How familiar are you with Joint Attention?



#	Field	Choice Count	
1	Not familiar at all	66.67%	20
2	Slightly familiar	26.67%	8
3	Moderately familiar	6.67%	2
4	Very familiar	0.00%	0
5	Extremely familiar	0.00%	0
			30

Showing rows 1 - 6 of 6

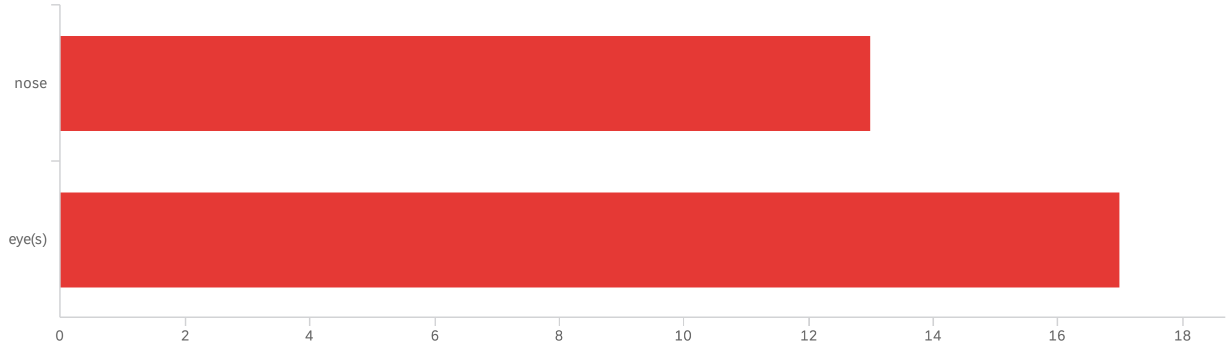
Post-Q2 - How well do you know how Joint Attention works?



#	Field	Choice Count	
1	Not well at all	63.33%	19
2	Slightly well	20.00%	6
3	Moderately well	10.00%	3
4	Very well	6.67%	2
5	Extremely well	0.00%	0
			30

Showing rows 1 - 6 of 6

Post-Q3 - Did you think Miro pointed with his nose or eye(s)



#	Field	Choice Count	
1	nose	43.33%	13
2	eye(s)	56.67%	17
			30

Showing rows 1 - 3 of 3

End of Report

Bibliography

[1] Michael Argyle. *Non-verbal communication in human social interaction.*, pages xiii, 443–xiii, 443. Cambridge U. Press, Oxford, England, 1972.

[2] David Mcneill. Hand and mind: What gestures reveal about thought. *Bibliovault OAI Repository, the University of Chicago Press*, 27, Jun 1994. doi: 10.2307/1576015.

[3] N.J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, Aug 2000. ISSN 0149-7634. doi: 10.1016/S0149-7634(00)00025-7.

[4] Zenzi Griffin and Kathryn Bock. What the eyes say about speaking. *Psychological science*, 11: 274–9, Aug 2000. doi: 10.1111/1467-9280.00255.

[5] Jean-David Boucher, Ugo Pattacini, Amelie Lelong, Gerrard Bailly, Frederic Elisei, Sascha Fagel, Peter Ford Dominey, and Jocelyne Ventre-Dominey. I reach faster when i see you look: Gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in Neurorobotics*, 6, 2012. ISSN 1662-5218. doi: 10.3389/fnbot.2012.00003. URL http://journal.frontiersin.org/article/10.3389/fnbot.2012.00003/abstract.

[6] Cesco Willemse and Agnieszka Wykowska. In natural interaction with embodied robots, we prefer it when they follow our gaze: a gaze-contingent mobile eyetracking study. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1771), Apr 2019. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2018.0036.

[7] Oriana Isabella Ferrari, Feiran Zhang, Ayrton A. Braam, Jules A. M. van Gurp, Frank Broz, and Emilia I. Barakova. Design of child-robot interactions for comfort and distraction from post-operative pain and distress. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, page 686–690, Stockholm Sweden, Mar 2023. ACM. ISBN 978-1-4503-9970-8. doi: 10.1145/3568294.3580174. URL https://dl.acm.org/doi/10.1145/3568294.3580174.

[8] Joshua Wainer, Ben Robins, Farshid Amirabdollahian, and Kerstin Dautenhahn. Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism. *Autonomous Mental Development, IEEE Transactions on*, 6:183–199, Sep 2014. doi: 10.1109/TAMD.2014.2303116.

[9] Minae Kwon, Malte F. Jung, and Ross A. Knepper. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 463–464, 2016. doi: 10.1109/HRI.2016.7451807.

[10] Barbora Siposova and Malinda Carpenter. A new look at joint attention and common knowledge. *Cognition*, 189:260–274, Aug 2019. ISSN 00100277. doi: 10.1016/j.cognition.2019.03.019.

[11] Tomoko Yonezawa, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. In *Proceedings of the ninth international conference on Multimodal interfaces - ICMI '07*, page 140, Nagoya, Aichi, Japan, 2007. ACM Press. ISBN 978-1-59593-817-6. doi: 10.1145/1322192.1322218.

[12] Olivia Barber, Eszter Somogyi, Anne E. McBride, and Leanne Proops. Children's evaluations of a therapy dog and biomimetic robot: Influences of animistic beliefs and social interaction. *International Journal of Social Robotics*, 13(6):1411–1425, Sep 2021. ISSN 1875-4791, 1875-4805. doi: 10.1007/s12369-020-00722-0.

[13] A Abubshait and E Wiese. You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human-robot interaction. *Frontiers in psychology*, 8, 2017. doi: 10.3389/fpsyg.2017.01393. URL `https://doi-org.tudelft.idm.oclc.org/10.3389/fpsyg.2017.01393`.

[14] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1):25, Mar 2017. ISSN 2163-0364. doi: 10.5898/JHRI.6.1.Admoni.

[15] Pauline Chevalier, Kyveli Kompatsiari, Francesca Ciardo, and Agnieszka Wykowska. Examining joint attention with the use of humanoid robots-a new approach to study fundamental mechanisms of social cognition. *Psychonomic Bulletin & Review*, 27(2):217–236, Apr 2020. ISSN 1531-5320. doi: 10.3758/s13423-019-01689-4.

[16] Sally J. Rogers. Mutual gaze. In Fred R. Volkmar, editor, *Encyclopedia of Autism Spectrum Disorders*, page 1966–1967. Springer New York, New York, NY, 2013. ISBN 978-1-4419-1697-6. doi: 10.1007/978-1-4419-1698-3_628. URL `http://link.springer.com/10.1007/978-1-4419-1698-3_628`.

[17] Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H. Johnson. Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences*, 99(14):9602–9605, 2002. doi: 10.1073/pnas.152159999. URL `https://www.pnas.org/doi/abs/10.1073/pnas.152159999`.

[18] Hironori Akechi and Harumi Kobayashi. Referential gaze and word mapping in autism spectrum disorders. In Vinood B. Patel, Victor R. Preedy, and Colin R. Martin, editors, *Comprehensive Guide to Autism*, pages 503–517. Springer New York, New York, NY, 2014. ISBN 978-1-4614-4788-7. doi: 10.1007/978-1-4614-4788-7_22. URL `https://doi.org/10.1007/978-1-4614-4788-7_22`.

[19] R Joanne Jao, Marybel Robledo, and Gedeon O Deák. The emergence of referential gaze and perspective-taking in infants.

[20] Maria Staudte and Matthew W. Crocker. Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition*, 120(2):268–291, Aug 2011. ISSN 00100277. doi: 10.1016/j.cognition.2011.05.005.

[21] Joy E. Hanna and Susan E. Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4):596–615, Nov 2007. ISSN 0749596X. doi: 10.1016/j.jml.2007.01.008.

[22] Douglas K. Candland. T. matsuzawa, m. tomonaga, m. tanaka (eds): Cognitive development in chimpanzees: Springer, tokyo, 2006, xvii + 522 pp., 239 figures, 26 in color, £54.00 (hardback). *International Journal of Primatology*, 28(4):965–967, Sep 2007. ISSN 0164-0291, 1573-8604. doi: 10.1007/s10764-007-9170-4.

[23] M. SCAIFE and J. S. BRUNER. The capacity for joint visual attention in the infant. *Nature*, 253 (5489):265–266, Jan 1975. ISSN 1476-4687. doi: 10.1038/253265a0.

[24] Tony Charman, Simon Baron-Cohen, John Swettenham, Gillian Baird, Antony Cox, and Auriol Drew. Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development*, 15(4):481–498, Oct 2000. ISSN 0885-2014. doi: 10.1016/S0885-2014(01)00037-5.

[25] Stephanie M. Carlson, Melissa A. Koenig, and Madeline B. Harms. Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):391–402, Jul 2013. ISSN 19395078. doi: 10.1002/wcs.1232.

[26] Susan R Leekam and Christopher A H Ramsden. Dyadic orienting and joint attention in preschool children with autism.

[27] Patrizia Piotti and Juliane Kaminski. Do dogs provide information helpfully? *PloS one*, 11(8): e0159797, 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0159797.

[28] Olivier Corneille, Sandie Mauduit, Rob. W. Holland, and Madelijn Strick. Liking products by the head of a dog: Perceived orientation of attention induces valence acquisition. *Journal of Experimental Social Psychology*, 45(1):234–237, Jan 2009. ISSN 0022-1031. doi: 10.1016/j.jesp.2008.07.004.

[29] Anna McPhee, Joseph Manzone, Matthew Ray, and Timothy Welsh. Timmy and lassie (and clyde?): Joint attention effects with humans, dogs, and orangutans. *PSYCHOMOTOR LEARNING ABSTRACTS*, 47(1), 2015.

[30] Bilge Mutlu, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction - HRI '09*, page 69, La Jolla, California, USA, 2009. ACM Press. ISBN 978-1-60558-404-1. doi: 10.1145/1514095.1514110. URL http://portal.acm.org/citation.cfm?doid=1514095.1514110.

[31] Zachary E. Warren, Zhi Zheng, Amy R. Swanson, Esubalew Bekele, Lian Zhang, Julie A. Crittendon, Amy F. Weitlauf, and Nilanjan Sarkar. Can robotic interaction improve joint attention skills? *Journal of Autism and Developmental Disorders*, 45(11):3726–3734, Nov 2015. ISSN 1573-3432. doi: 10.1007/s10803-013-1918-4.

[32] Deirdre E. Logan, Cynthia Breazeal, Matthew S. Goodwin, Sooyeon Jeong, Brianna O'Connell, Duncan Smith-Freedman, James Heathers, and Peter Weinstock. Social robots for hospitalized children. *Pediatrics*, 144(1), Jul 2019. ISSN 0031-4005. doi: 10.1542/peds.2018-1511. URL https://doi.org/10.1542/peds.2018-1511.

[33] Stela H. Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E. Young. Poor thing! would you feel sorry for a simulated robot? a comparison of empathy toward a physical and a simulated robot. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 125–132, 2015.

[34] Joshua Wainer, David J. Feil-Seifer, Dylan A. Shell, and Maja J. Mataric. Embodiment and human-robot interaction: A task-based perspective. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 872–877, 2007. doi: 10.1109/ROMAN.2007.4415207.

[35] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies*, 64(10):962–973, Oct 2006. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2006.05.002.

[36] C.D. Kidd and C. Breazeal. Effect of a robot on user perceptions. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 4, pages 3559–3564 vol.4, 2004. doi: 10.1109/IROS.2004.1389967.

[37] Abdulaziz Abubshait, Patrick Weis, and Eva Wiese. Effects of embodiment on social attention mechanisms in human-robot interaction. *Frontiers in Human Neuroscience*, 12, 2018. doi: 10.3389/conf.fnhum.2018.227.00080.

[38] Molly C. Martini, Christian A. Gonzalez, and Eva Wiese. Seeing minds in others – can agents with robotic appearance have human-like preferences? *PLOS ONE*, 11(1):e0146310, Jan 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0146310.

[39] Friederike A. Eyssel and Michaela Pfundmair. Predictors of psychological anthropomorphization, mind perception, and the fulfillment of social needs: A case study with a zoomorphic robot. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 827–832, 2015. doi: 10.1109/ROMAN.2015.7333647.

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

[41] Dingjun Li, Patrick P.L. Rau, and Ye Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2:175 – 186, May 2010. doi: 10.1007/s12369-010-0056-9.

[42] Sara Kiesler and Jennifer Goetz. Mental models of robotic assistants. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, page 576–577, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134541. doi: 10.1145/506443.506491. URL https://doi-org.tudelft.idm.oclc.org/10.1145/506443.506491.

[43] Aaron Powers and Sara Kiesler. The advisor robot: Tracing people's mental model from a robot's physical attributes. In *HRI 2006: Proceedings of the 2006 ACM Conference on Human-Robot Interaction*, volume 2006, page 218–225, Jan 2006. doi: 10.1145/1121241.1121280.

[44] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Suzana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int J of Soc Robotics*, 1:71–81, 2009. doi: 10.1007/s12369-008-0001-3.

[45] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. Artificial social agent questionnaire instrument, 2023. URL https://data.4tu.nl/datasets/d8e2c534-d192-4411-a1e2-5aeac97f0165/6.

[46] Maria T. H. van Otterdijk, Manon W. P. de Korte, Iris van den Berk-Smeekens, Jorien Hendrix, Martine van Dongen-Boomsma, Jenny C. den Boer, Jan K. Buitelaar, Tino Lourens, Jeffrey C. Glennon, Wouter G. Staal, and Emilia I. Barakova. The effects of long-term child–robot interaction on the attention and the engagement of children with autism. *Robotics*, 9(4), 2020. ISSN 2218-6581. doi: 10.3390/robotics9040079.

[47] Jaime Banks, Kevin Koban, and Philippe Chauveau. Forms and frames: Mind, morality, and trust in robots across prototypical interactions. *Human-Machine Communication*, 2:81–103, January 2021. ISSN 2638-6038, 2638-602X. doi: 10.30658/hmc.2.4.

[48] Jacob Cohen. Quantitative methods in psychology a power primer. 1992. URL https://api.semanticscholar.org/CorpusID:14411587.

[49] URL https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.

[50] Björn Walther. Friedman test / friedman test - calculate required sample size with g*power, Aug 2023. URL https://www.youtube.com/watch?v=QKugf-lXqDk.

[51] Mark Williamson. Sample size calculation with gpower.

[52] Kyveli Kompatsiari, Vadim Tikhanoff, Francesca Ciardo, Giorgio Metta, and Agnieszka Wykowska. The importance of mutual gaze in human-robot interaction. In Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He, editors, *Social Robotics*, volume 10652 of *Lecture Notes in Computer Science*, page 443–452. Springer International Publishing, Cham, 2017. ISBN 978-3-319-70021-2. doi: 10.1007/978-3-319-70022-9_44. URL http://link.springer.com/10.1007/978-3-319-70022-9_44.

[53] Nov 2023. URL https://www.qualtrics.com/.

[54] . URL https://consequentialrobotics.com/.

[55] . URL https://www.miro-e.com/our-story.

[56] . URL https://www.miro-e.com/faq.

[57] . URL https://www.miro-e.com/research-applications.

[58] Ikuhisa Mitsugami, Norimichi Ukita, and Masatsugu Kidode. Robot navigation by eye pointing. *Lecture Notes in Computer Science*, 3711:256–267, 01 2005.

[59] . URL https://pupil-labs.com/products/core/.

[60] Amanda Dattalo. Ros introduction, Aug 2018. URL https://wiki.ros.org/ROS/Introduction.

[61] . URL http://labs.consequentialrobotics.com/miro-e/software/.

[62] . URL https://docs.pupil-labs.com/core/software/pupil-capture/.

[63] Sabbir Ahmed. Real-time head pose estimation with opencv and dlib, Jun 2020. URL https://medium.com/analytics-vidhya/real-time-head-pose-estimation-with-opencv-and-dlib-e8dc10d62078.

[64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[65] Nicholas Renotte. Build a deep face detection model with python and tensorflow | full course, May 2022. URL https://www.youtube.com/watch?v=N_W4EYtsa10.

[66] Kentaro Wada. Labelme: Image Polygonal Annotation with Python. URL https://github.com/wkentaro/labelme.

[67] E Schmider, M Ziegler, E Danay, and M Beyer, L & Bühner. Is it really robust? reinvestigating the robustness of anova against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6(4):147–151, 2010. doi: 10.1027/1614-2241/a000016.

[68] URL https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php#:~:text=The%20Friedman%20test%20is%20the,variable%20being%20measured%20is%20ordinal.

[69] Jacob Cohen. Statistical power analysis for the behavioral sciences. 1988. doi: 10.4324/9780203771587.

[70] URL https://www.rdocumentation.org/packages/rstatix/versions/0.7.2/topics/identify_outliers.