Designing Joint Attention Systems for robots that assist children with autism spectrum disorders

LEONOR FERMOSELLE
Designing joint attention systems for robots that assist children with autism spectrum disorders

LEONOR FERMOSELLE

Master in Systems, Control and Robotics
Date: June 19, 2018
Supervisor: Joakim Gustafson and Isabel Trancoso
Examiner: Jonas Beskow
Swedish title: Utforma gemensamma uppmärksamhetssystem för robotar som hjälper barn med autismspektrumstörningar
School of Electrical Engineering and Computer Science
Abstract

Joint attention behaviours play a central role in natural and believable human-robot interactions. This research presents the design decisions of a semi-autonomous joint attention robotic system, together with the evaluation of its effectiveness and perceived social presence across different cognitive ability groups. For this purpose, two different studies were carried out: first with adults, and then with children between 10 and 12 years-old.

The overall results for both studies reflect a system that is perceived as socially present and engaging which can successfully establish joint attention with the participants. When comparing the performance results between the two groups, children achieved higher joint attention scores and reported a higher level of enjoyment and helpfulness in the interaction.

Furthermore, a detailed literature review on robot-assisted therapies for children with autism spectrum disorders is presented, focusing on the development of joint attention skills. The children’s positive interaction results from the studies, together with state-of-the-art research therapies and the input from an autism therapist, guided the author to elaborate some design guidelines for a robotic system to assist in joint attention focused autism therapies.
Sammanfattning


De övergripande resultaten för båda studierna visar att vi byggt ett system som uppfattas som socialt närvarande och engagerande, och som framgångsrikt kan skapa gemensam uppmärksamhet med deltagarna. När man jämför resultaten mellan de två grupperna, finner man att barn gav högre gemensam uppmärksamhetsresultat de rapporterade att de fick mer hjälp av och tyckte bättre om roboten som använda det utvecklade systemet för joint attention.

Vidare presenteras en detaljerad litteraturstudie om robotassisterade terapier för barn med autismspektrumsjukdomar, med fokus på utveckling av gemensamma uppmärksamhetsförmågor. Barnens positiva interaktioner med en robot som hade det utvecklade joint attentionsystemet, tillsammans med litteraturstudier om forskningsterapier och en intervju med en autismterapeut, vägledde författaren att utarbeta riktlinjer för hur man ska designa sociala robotar som har till syfte att användas vid terapi som syftar till att förbättra autistiska barns förmåga att kommunicera icke-verbalt och förstå vad den de talar med fokuserar på.
Acknowledgements

First of all, I would like to thank my parents for their support on my academic journey and particularly to my mother, for the wise advises and for being an example of strength.

I would also like to thank to my supervisors Joakim Gustafson and Isabel Trancoso, for the opportunity to participate in such a wonderful project.

I owe special thanks to André Pereira, for all the valuable discussions and guidance throughout this research, and for the opportunity to work together.

Finally, my most affective thanks to Miguel Moreira, among other things, for the visits and for always listening to my worries. Even though from a distance, you were always there.
## List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABA</td>
<td>Applied Behaviour Analysis</td>
</tr>
<tr>
<td>ADOS</td>
<td>Autism Diagnostic Observation Schedule</td>
</tr>
<tr>
<td>ASD</td>
<td>Autism Spectrum Disorders</td>
</tr>
<tr>
<td>EBP</td>
<td>Evidence Based Practice</td>
</tr>
<tr>
<td>ESCS</td>
<td>Early Social Communication Scale</td>
</tr>
<tr>
<td>HRI</td>
<td>Human-Robot Interaction</td>
</tr>
<tr>
<td>IJA</td>
<td>Initiating Joint Attention</td>
</tr>
<tr>
<td>JA</td>
<td>Joint attention</td>
</tr>
<tr>
<td>KTH</td>
<td>Kungliga Tekniska Högskolan</td>
</tr>
<tr>
<td>RJA</td>
<td>Responding to Joint attention</td>
</tr>
<tr>
<td>SAR</td>
<td>Socially Assistive Robotics</td>
</tr>
<tr>
<td>TD</td>
<td>Typically Developing</td>
</tr>
<tr>
<td>TMH</td>
<td>Tal, Musik och Hörsel</td>
</tr>
<tr>
<td>VPT</td>
<td>Visual Perspective Taking</td>
</tr>
<tr>
<td>WoZ</td>
<td>Wizard of Oz</td>
</tr>
</tbody>
</table>
List of Figures

3.1 Robots used in ASD therapy .................................. 21
4.1 Front view of the main MagPuzzle setup ................. 26
4.2 Wizard of Oz interface screenshot ............................ 29
6.1 Average rate of time of the robot’s gaze distribution with adults. .............................................. 39
6.2 Frequency of third party observers’ opinions ............. 40
6.3 Average percentage of time of mismatched and shared attention with adults .............................. 42
6.4 Mean values of the final questionnaire measures with adults ...................................................... 43
6.5 Mean values of the six dimensions of the Social Presence measure [28] in each direction ............. 44
6.6 Average percentage of mismatched and shared attention with children .............................. 46
6.7 Mean values of the final questionnaire measures with children ............................................... 47
6.8 Average percentage of mismatched and shared attention for adults and children in full joint attention condition ......................................................... 50
6.9 Mean values of the final questionnaire measures for adults and children in full joint attention condition ................................................................. 52
6.10 Mean values of the six dimensions of Social Presence measure for adults and children in full joint attention condition ............................................ 53
List of Tables

3.1 Compilation of JA related measures ................. 14
3.2 Robot design features [29] .......................... 18
3.3 Compilation of relevant studies for developing JA skills . 22

4.1 Illustrating examples of dialogue acts behaviours .... 27
4.2 List of gaze shift trigger events with respective priority and enabling condition .......................... 32

7.1 Illustrating examples of ABR therapy for training joint attention ........................................ 57
# Contents

1 Introduction 1

1.1 Motivation 1

1.2 Objective 2

1.3 Outline 2

2 Background 4

2.1 Importance of joint attention skills 4

2.2 Defining Autism Spectrum Disorders 5

2.3 Prevalence of Autism Spectrum Disorders 7

2.4 Importance of ASD therapy 8

3 Related Work 10

3.1 Joint Attention Measures 11

3.2 Design decisions for robot-assisted therapy 15

3.2.1 Important remarks 19

3.3 HRI therapy for developing JA in children with ASD 20

4 Joint Attention Task 24

4.1 Task Description and Hardware 24

4.2 Dialogue Acts and Wizard Interface 27

4.3 Multimodal Perception 29

4.4 Autonomous joint attention system 31

5 Methodology 33

5.1 Goal of the studies 33

5.2 Participants 34

5.3 Procedure 34

5.4 Manipulation 35

5.5 Measures 35
6 Results and Discussion

6.1 Adult’s performance ........................................ 41
   6.1.1 Objective Results ................................... 41
   6.1.2 Subjective Results ................................ 42

6.2 Children’s Performance .................................. 45
   6.2.1 Objective Results ................................... 45
   6.2.2 Subjective Results ................................ 46

6.3 Comparison and Discussion ............................... 48
   6.3.1 Objective Results ................................... 48
   6.3.2 Subjective Results ................................ 50

7 Guidelines for robotic systems to assist children with ASD 54

7.1 Interview with Autism Expert ............................. 56
7.2 Extending Design Guidelines .............................. 58

8 Conclusion ..................................................... 60

8.1 Final Discussion ........................................... 62
8.2 Limitations and Future Work ............................. 63

Bibliography .................................................... 65

A Publication .................................................... 71

A.1 ICMI 2018: The 20th ACM International Conference on
     Multimodal Interaction .................................... 71
Chapter 1

Introduction

1.1 Motivation

Intelligent robots are becoming part of our daily lives. The development of autonomous robotic technologies raises new application domains for this field such as health care, transportation or work on production lines.

An emerging field of autonomous robots are social robots. As the name suggests, social robots [1] are especially designed to interact with people, helping them to perform tasks in several environments, providing company to the elderly [2] and even encouraging the development of social behaviours [3]. For this, social robots have to recognize and respond to human social cues through appropriate behaviours, such as verbal communication or non-verbal communication, as the expression of facial emotions, gestures, responding to joint attention, but also to elicit responses from the human side in order to establish a social connection. There has been a growing interest in designing interactive robots that humans can naturally interact with, and, therefore, researchers have focused on how to improve the interaction to make it more believable and natural [4].

In health care, social robots have been known to positively impact the therapy sessions of children with autism spectrum disorders, who face moderate to severe social impairments throughout their lives and greatly benefit from treatment. As a matter of fact, social robots have been shown to trigger a high degree of motivation and proactive behaviours in children with autism, even between subjects that are un-
likely or unwilling to interact with human therapists or care-givers [5] which reinforces the role of social robots as a promising tool for autism therapies.

Among the social impairments children with autism spectrum disorders face, the lack of joint attention skills is considered to be of great importance, given that it is the basis of language and social-cognitive development. Impairments in this skill can be a reliable diagnosis factor for autism from early on. It is important to acknowledge that, over the last decades, there has been a markedly increasing prevalence of autism spectrum disorders [6] and therefore, it is necessary to provide alternative or assistive treatments for this pathology through the development of technology, given the high costs and accessibility issues traditional therapies have.

For these reasons, a considerable amount of research has focused on assisting children with ASD to develop joint attention skills through human-robot interaction, using friendly social robots.

1.2 Objective

The goal of this thesis is to evaluate the outcomes and the perceived effects of the design decisions in an autonomous joint attention system, throughout different age groups with different cognitive abilities, both adults and children. Furthermore, special design guidelines for joint attention systems are investigated for assisting in the therapy of a particular demographic group with impairments on this area, children with ASD. Hence, this thesis has the following research question:

*How to design and evaluate autonomous joint attention mechanisms for different cognitive ability groups*

1.3 Outline

This thesis is organized in nine different chapters:

Chapter 2 (Background) presents a brief characterization of joint attention skills and their importance in human development, autism spectrum disorders and its prevalence rates, followed an analysis of the importance of therapy.
Chapter 3 (Related Work) presents a detailed literature review of recent robot-assisted therapies for children with ASD, focusing on common design features followed by most systems. Then, an analysis of human-robot therapies for developing joint attention skills in children with ASD is presented together with a compilation of their findings, and finally, a review of some important joint attention qualitative and quantitative measures is elaborated.

Chapter 4 (Joint Attention Task) provides a detailed description of the developed multimodal, autonomous joint attention system, which establishes joint attention through the interaction with a robotic head, by speech recognition and gaze estimation in the context of a physical game: the MagPuzzle task.

Chapter 5 (Methodology) describes the methodology followed by two different studies, carried out with adults and children, using the MagPuzzle task exposed on the previous chapter.

Chapter 6 (Results and Discussion) starts with a demonstration of the different behaviours the system has in the two manipulation conditions and its perceived effects by a third party observer study. This is followed by an analysis of the quantitative and qualitative results obtained in each of the two studies, with adults and children, followed by another analysis of performance differences.

Chapter 7 (Guidelines for robotic systems to assist children with ASD) presents some design guidelines for robotic systems to assist in therapy of children with ASD for developing joint attention skills, elaborated with the analysis of the results, literature study and together with the input from an interview with an autism therapy expert.

Chapter 8 (Conclusion), as the name suggests, closes this thesis with an overview of the contributions made in this research and relevant findings, together with a final discussion and an analysis of limitations and future work.
Chapter 2

Background

This chapter provides some definitions and context about joint attention skills, their importance in human-robot interactions, a characterization of autism spectrum disorders (ASD) and the importance of therapy. With this, the author hopes to provide a better understanding of these concept’s and disorder’s impacts, and to facilitate the reading of this thesis, where these terms will be used extensively.

2.1 Importance of joint attention skills

Joint attention is a social process that involves sharing focus with another person, object or event [7]. It is a nonverbal social-communication skill and it can have several forms such as gaze alternating, which can involve checking someone’s face repeatedly during an interaction event or sharing focus in ambiguous situations, gestures (pointing, showing) and even coordinating forms [8]. For example, when an individual alerts another to a target by means of eye-gazing, pointing or other indirect verbal indications and the other person looks back to her after looking at the object. However, it is worth highlighting that it is the social interaction function inherent in joint attention prompts that makes it more than just a repertoire of gazes, gestures and vocalizations.

Joint attention capabilities are the initial means of communication for infants, allowing them to start sharing experiences and to comprehend shared meanings [9]. It provides a context for experiencing and obtaining knowledge about the world and about others and opens op-
portunities for developing communication with adults and other children [9]. As a matter of fact, there is a considerable amount of literature that supports the view of joint attention skills as a pivotal skill [10, 11, 8] or in other words, a behaviour that when subject to intervention will result in positive collateral changes of other untreated behaviours, in this case language and social-cognitive development.

However, joint attention has been shown to be not only important for human-human interaction but also for human-robot interaction. It can help to make cooperation more efficient and support disambiguation in instances of uncertainty [12]. By endowing robots with subtle joint attention mechanisms together with other verbal behaviours, it can create a more engaging interaction and be perceived by the users as more socially present. These are important factors for achieving believable long-term interactions.

2.2 Defining Autism Spectrum Disorders

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [13] and the International Statistical Classification of Diseases and Related Health Problems (ICD10) [14], autism spectrum disorders are classified as pervasive developmental disorders and are defined by the presence of three essential features in an individual: marked impairments in social interactions, marked impairments in communication and very restricted and repetitive behaviour patterns and interests.

The diagnostic criteria for autism spectrum disorders in [13] states that such condition can be spotted through several different behaviours and personality traits, but for being considered an autism spectrum disorder there must be at least two symptoms related to social interactions impairments, at least one in atypical communication patterns and one symptom in repetitive behaviours.

Impairments in social interactions include symptoms covering accentuated deficits in nonverbal behaviours such as facial expressions, eye-contact or joint attention, failure to develop peer relationships appropriate to developmental level, lack of spontaneous interest in sharing enjoyment or interests and lack of social reciprocity.

Moreover, the atypical communication patterns are reflected in the
lack of, or accentuated delay in developing spoken language, not compensated with alternative gestures or mime, and in the ones that have developed spoken communication they show marked impairments in the ability to initiate or maintain a conversation, repetitive and stereotyped use of language, and lack of varied make-believe play appropriate to developmental level.

In what concerns the repetitive behaviours and interests, they can be manifested by a significant preoccupation with one or more stereotyped and restricted patterns of interest that are abnormal whether in intensity or focus (for example an interest in massive amounts of meteorology facts), an apparent inflexibility in adherence to changes in routines or rituals, stereotyped and repetitive motor movements such as hand or finger clapping or whole-body movements, and even a persistent preoccupation with specific parts of objects.

It is known that deficits in joint attention capabilities are a reliable factor for ASD diagnosis from early on, often evident before 1 year of age and before any diagnosis has been made [15]. This impairment is an early social disturbance, related with neurological, cognitive and affective processes that may play a role in autism [10]. Therefore, it is crucial that primary behavioral interventions focus on developing this skill in young children with ASD [8] to enable potential developments of social-cognitive skills and spoken language.

Typical ASD joint attention impairments are reflected by deficits in both responding to others’ joint attention directives (RJA) and in initiating joint attention (IJA). Yet, some studies indicate that joint attention impairments in children with ASD change over time: although initiating joint attention bids remains impaired, some children that show higher cognitive development can start to demonstrate ability to respond to others’ joint attention directives [8, 15]. Bearing in mind the differential development between responding to and initiating joint attention directives, it is possible that the two skills have different functions and that only initiating JA bids has a social purpose [8], being the one that reflects typical social impairments from ASD.

Apart from the formal symptoms described in [13], another typical impairment present in ASD is the lack of Visual Perspective Taking (VPT) [16]. This skill (VPT) refers to the capacity of understanding that others have thoughts, feelings and perspectives different from ours,
which allows to predict other people’s behaviours and thoughts based on what we think they might be thinking [16]. This ability typically develops and becomes consistent between 4 and 6 years of age in typically developing children. However, children with ASD, specifically, fail to employ VPT as Baron-Cohen et al. reported in [17], which naturally affects social interactions and communication.

The cause for ASD is unknown and there is no cure, but research has shown that its symptoms can be improved through proper treatment, particularly when detected in early childhood. Autism disorders are usually diagnosed in a developing phase of an individual, more specifically in his/her childhood with more accentuated symptoms over time. In fact, by definition in [13] the onset of ASD is prior to 3 years old although manifestations of the disorder before 2 years old are more difficult to define and to diagnose than those after that age.

The disorder follows a continuous progression over time where the development of spoken language and overall intellectual level are the ultimate factors to the diagnosis. According to [13], about one-third of ASD cases follow to live with some degree of independence and the highest functioning adults with ASD demonstrate impairments in social interactions and very restricted interests.

2.3 Prevalence of Autism Spectrum Disorders

Epidemiological surveys concerned with Autism Spectrum Disorders started mid-1960s in England [6]. However, nowadays there is no consensus about the prevalence of ASD in the population.

According to DSM-IV [13], epidemiological studies suggest that ASD affects 2-5 people in 10000. In Fombonne epidemiology of autism review [6], which surveyed more than 4 million subjects across the globe in several studies 1533 subjects were identified as having autism and the estimated median prevalence was 5.2/10000 across surveys. The average male/female prevalence rate was 3.8/1, depending on the absence or presence of mental retardation, and factors such as social class or immigrant status had no apparent correlation or causation with autism disorders.
Nonetheless, over the last decades there seems to be a markedly increase in the prevalence of this condition [6, 18]. In [6] a correlation between study’s year of publication and prevalence rate was found with statistical significance (p-value < 0.01), although the extent to which the increasing prevalence pattern can be attributed to a combination of shifting diagnosis tendencies and increased ASD awareness, or changes in true risk is unclear.

2.4 Importance of ASD therapy

Given the increasing prevalence tendency of ASD and the lifelong effects and struggle in life that children with this disorder face, an effective and customized therapy plays a fundamental role in the child’s understanding of the world and others, in developing healthy interactions with other people and ultimately in gaining some independence. For instance, Mundy and Crowson in [10] state that young children who are presented with intensive behavioral interventions during a considerable amount of time, demonstrate substantial IQ gains.

Hence, an early diagnose of the disorder and an individual assessment of the child’s symptoms for a successful behavioral intervention and possible functional improvements is considered by some a public health emergency [18].

For early diagnose, the American Academy of Neurology recommends that children who fail typical routine developmental exams to be screened with either the Modified Checklist for Autism in Toddlers (M-CHAT) or the Autism Screening Questionnaire [18]. Moreover, there are relevant findings in the study carried out in [19] that point out the possibility for diagnose of ASD through the analysis of movement patterns as early as 4-6 months old. This study surveyed 17 children with ASD, diagnosed after 3 years old by conventional methods, and analyzed videos of their infancy provided by the parents, finding specific movement disturbances from a very early age in these children, which indicates the need for the development of therapy and intervention from the first few months of life in autism.

Behavioural intervention tools for mitigating ASD impairments focus on overcoming the three characteristic features of ASD, social interaction impairments, communication impairments and restricted and
repetitive movement patterns and interests. For example, applied behaviour analysis (ABA) is a scientifically validated approach for understanding behaviours and how they are affected by the environment, and includes treatment techniques to bring out positive meaningful changes in behaviour [20].

ABA techniques focus on how learning takes place, for instance through positive reinforcement (when a behaviour is followed by some reward it is more likely to be repeated), and in what concerns autism disorders these techniques can foster basic skills such as looking, hearing, imitating and more complex skills like reading and understanding another person’s perspective. There are several completed studies with age groups ranging from preschoolers to adults [21] that prove that ABA techniques can produce significant improvements in communication, social interaction, self-care, school and employment as well as increased participation in family and community activities.
Chapter 3

Related Work

The literature investigated in this chapter covers some state-of-art measures of joint attention and provides some insight on current technology-based therapies and interventions for children with ASD, focusing on the robotic-systems’ design decisions and on the development of joint attention skills.

The development of technology targeted towards ASD therapy is extremely relevant. Traditional behavioural sessions or interventions are not accessible for the whole ASD population due to the lack of trained therapists and specialists and to the associated costs of interventions, hence technology promises an alternative assistive treatment, increases the accessibility of intervention, reduces assessment efforts and the cost of therapy. This promotes intervention and ultimately the generalization the trained skills to new environments beyond intervention [22].

The literature review presented is organized based on the following key points: (1) Joint attention measures; (2) Revisions of design decisions of robot-assisted therapy systems and (3) Human-robot interaction therapies targeted towards developing joint attention skills. Short summaries or compilations of the most important remarks are presented in the end of each section.
3.1 Joint Attention Measures

Using social robots to quantitatively measure social behaviour cues, for example through gaze direction, position tracking or vocal prosody, constitutes an important tool to diagnose, treat and better understand ASD [23]. In this field, joint attention measures appear to be among the more powerful early diagnostic tools for autism pathology [15].

There are several effective measures for joint attention in children, like the Early Social Communication Scale (ESCS) [24] and the Autism Diagnostic Observation Schedule (ADOS). These two measures are appropriate for children up to 30 months of developmental age (ESCS) and those with two-word phrase speech (ADOS).

The original ESCS measure [25] is a videotaped observation measure that takes 15 to 25 minutes to administer. It consists of a set of 25 semi-structured eliciting situations to encourage interaction between an adult tester and the child, where 110 different child behaviours occurrences were found and coded according to developmental stage, communicative goal and whether the child initiated the interaction or responded to the tester’s bid. The reduced version of the original ESCS [24] was designed as a more practical research and clinical tool, that emphasizes the frequency of the data (number of prompts and targets), dividing the complexity of the behaviours in lower vs. higher level behaviours. The behaviours of interest in this measure have three dimensions: Joint attention behaviours (IJA and RJA), Behavioural requests (initiating and responding) and Social interaction (initiating and responding). IJA refers to the frequency with which the child uses eye contact, pointing and showing to initiate shared attention to objects or events and RJA refers to the child’s skill in following the tester’s line of regard and pointing gestures. Detailed scoring tables for the frequency of each type of behaviour according to the situation are available in [24].

However, those measures do not seem appropriate for assessing or measuring joint attention in school-age or adolescence, because they don’t capture the subtlety and variations in the development of JA for high functioning, older individuals on the spectrum.

In [7], the authors developed a measure of response to JA (RJA) appropriate for subjects with ages between 7 and 17 years-old, which
yields a more continuous, graded measure of JA skills. This JA scale consists of a series of six prompts designed to elicit a joint attention response in which the tested subject has to disengage from its current focus of attention. The prompts were divided in two types: (1) whether experimenter’s prompt was verbal, and (2) the number of attentional shifts required for a "correct response". The scores for responding to JA were attributed according to 4 categories: (1) engaging in triadic attention; (2) looking at examiner’s face; (3) making eye contact, and (4) offering a spontaneous, relevant verbalization, in which higher scores translate into more social responses.

This experimental JA measure [7] was assessed with a group of individuals with ASD (n=18) and a group of TD individuals (n=24). The final results reported showed that the total scores for RJA differed significantly for ASD and TD groups (p-value<0.01) and had great variability within each group. Furthermore, significant correlations were found between better performance with this experimental JA measure and better performance on measures of receptive language, but not with social relatedness measures such as ADOS. Additionally, this measure also reported a range of scores for TD participants according to their age, suggesting that JA varies in adolescence and that this measure is sensitive enough to capture subtle differences.

In more recent research, Séverin Lemaignan et al. [26] proposed in 2016 an indirect new measure for engagement entitled with-me-ness. Engagement is a broad concept, with many possible interpretations which makes it difficult to operationalize in the context of human-robot interaction. The introduction of this more specific concept with-me-ness, aims to quantify the extent to which the human is "with-me", the robot, during an interactive task (on-line). The goal is achieved by comparing the focus of the human attention, estimated in real-time by the robot through a fast head-pose estimation, with the expected (a priori) targets of attention elicited by the robot. In other words: "how well the child is following the robot’s expectations".

The concept of with-me-ness was borrowed from Sharma et al. [27] as a gaze-measure of students’ attention, and was adapted for the context of human-robot interaction by S. Lemaignan et al. [26] who defined conceptual with-me-ness as the normalized ratio of time that the participant focuses his/her attention on the targets elicited by the
robot for the current task (algorithm available in [26]). By analyzing the results from this measure, it is possible to obtain a visual graph of the child’s attention distribution to each of the targets over time, as well as an accurate picture of the overall turn-taking, as perceived by the robot. Thus, with-me-ness measure complies with the notion of joint attention in the sense that it includes information about shared focus of attention between the participant and the robot.

The experimental validation of this measure was carried out with 6 TD children in a face-to-face interaction with an autonomous robot achieving accurate enough results for reliable computation of a metric of attention focus which can be applied, for example, as a real-time feedback to a robot controller to build more adaptive interactive behaviours.

Also related to joint attention is the social presence concept. Better simulations of joint attention in an artificial agent might result in an improved sense of social presence from the user’s perspective. Hence, a measure of social presence is useful to operationalize the effects of joint attention on the user’s side, which makes these concepts interconnected.

In [28], F. Biocca et al. defined social presence in a mutual interaction with a specific entity as the degree of initial awareness, allocated attention, capacity for both content and affective comprehension and capacity for both affective and behavioural interdependence with the given entity. This concept is therefore broader than joint attention, but it includes behavioural features characteristic of JA cues that might be good evaluation components for this social behaviour.

For this purpose, [28] developed a measure of social presence entitled networked minds. This measure decomposes social presence in six dimensions: Co-presence, attentional allocation, perceived message understanding, perceived affective understanding, perceived affective interdependence and perceived behavioural interdependence. The measure itself consists of 36 items that reflect the six dimensions of social presence, where half of the items are repeated questions but in the observer’s perspective of the other’s response, for symmetry purposes. These items are answered following the interaction session. A detailed list of the items composing this measure is available in [28].

The validation of this measure was carried out through a study
with n=240 participants with three conditions: face-to-face interactions, mediated interaction via text-based media and mediated interaction via video-conferencing media. The results obtained point out that the measure was able to distinguish between social presence experience of face-to-face interaction and mediated interaction in at least four of the different dimensions, though it may not be sensitive enough to detect differences in social presence across different media.

Table 3.1 is presented to summarize the main features of the joint attention related measures analyzed in this section.

Table 3.1: Compilation of JA related measures

<table>
<thead>
<tr>
<th>Measure title</th>
<th>Method</th>
<th>Measurement type</th>
<th>Target audience</th>
<th>Relevant features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESCS [24]</td>
<td>Set of 25 semi-structured situations eliciting interaction between adult tester and child</td>
<td>Frequency of responses, gazes and initiatives of interaction</td>
<td>Children up to 30 months</td>
<td>Measures 3 dimensions of social behaviours: joint attention, behavioural requests and social interaction</td>
</tr>
<tr>
<td>With-me-ness</td>
<td>Track participant’s gaze with a fast head-pose estimator and compare with predefined desired targets of attention throughout time</td>
<td>Temporal measure of participant’s gaze</td>
<td>General public</td>
<td>Possible to obtain visual graphic of targets of attention over time and overall turn-taking of attention</td>
</tr>
<tr>
<td>RJA in adolescents [7]</td>
<td>Six examiner-initiated prompts to elicit a JA response, requiring the participant to change focus of attention</td>
<td>Observational behaviour measure</td>
<td>Individuals ages 7 to 17</td>
<td>Capture subtleties in JA development through adolescence and differences between TD and ASD adolescents</td>
</tr>
<tr>
<td>Networked Minds [28]</td>
<td>Answering questionnaire with 36 items following the interaction session</td>
<td>Questionnaire based social presence measure</td>
<td>General public with capacity for answering questionnaires</td>
<td>Measures 6 dimensions of social presence</td>
</tr>
</tbody>
</table>
3.2 Design decisions for robot-assisted therapy

Socially assistive Robots (SAR) aim to provide assistance to human users through social interactions [1] by recognizing and responding to human social cues through appropriate behaviours, such as communication, expressing and perceiving emotions, maintaining social relationships, or developing social competences.

In the context of ASD therapy, social robots are being used as tools to assist in the diagnosis of autism disorder, to teach skills to children with autism, to play with them and to elicit certain behaviours from them [29]. Indeed, robots have been shown to trigger a high degree of motivation and proactive behaviours in children with ASD, even between subjects that are unlikely or unwilling to interact with human therapists or care-givers [5, 23, 30], which reinforces the role of social robots as a promising tool for ASD and other disorders’ therapies.

In recent past, this field has faced a rapid growth with innumerous research initiatives being carried out worldwide by agencies and universities with the purpose of developing social robots and conducting clinical tests on children with autism. Naturally, there is a great number of studies proposing and reviewing ASD therapy approaches using robots with different appearances, behaviours, activities that they can do, and experiments with distinct perspectives, criteria and organization.

The use of robots in autism therapy must balance a goal-oriented treatment with an engaging, yet nonthreatening, productive interaction. Scasselati et al. [31] presented in 2012 a review of the previous decade of research in this field, focusing on (1) Robot design, (2) Human-Robot interaction and (3) System evaluation.

In what concerns robot design, the authors discuss different levels of anthropomorphism for robots used in ASD therapy, contrasting humanoid with non-humanoid robots, and whether robot anthropomorphism had an impact in the outcomes of ASD therapy, such as generalization of the skills learned in therapy and engagement levels. The authors point out that humanoid robots facilitate the goal of generalizing the skills learned in therapy to a reality situation, whereas the
maximum engagement was found with non-humanoid robots. Furthermore, animal-like or cartoon-like robots with oversized and exaggerated primary features might help to focus the attention on specific social cues that are necessary for training a skill. Ricks and Colton’s study [32] also supports these findings.

Human-robot interaction is addressed from two perspectives, the targeted behaviour and the robot’s role in the interaction, and completed with an analysis of robot’s autonomy. The targeted behaviour refers to the goal of an interaction between human user and robot which might be to elicit joint attention between the user and the robot, to mediate sharing turn-taking between user and others, or to encourage imitative behaviours. At the same time, the role of the robot may differ depending on the situation, it can act as a teacher in an authoritative role, as a toy intended to mediate behaviours, or as a proxy for the user to allow him or her to express emotions or goals. Concerning robot’s autonomy, the authors highlight the need to achieve an autonomous robotic system for ASD therapies given the limitations of operator-controlled methods, such as the Wizard of Oz scheme.

Finally, regarding system evaluation, Scasselatti et al. analyzes robot-assisted ASD therapies based the number of interactions required (single or repeated), sample sizes, structure of interaction (guided by the presence of a therapist or relative free-form interaction) and data collection analysis.

Pennisi et al.[30] presented in 2015 a systematic review of social robots used in ASD therapies among 29 different studies, focusing the review in a "skills" perspective. The major contribution of [30] is related to the formulation of 10 questions regarding the essential roles and benefits of robots for ASD treatments, after grouping the elected studies according to the tested ability (social behaviour, joint attention, imitation and language). The authors’ findings were reported in the light of these questions and the overall results point out positive observations in prosocial behaviours, maintenance of attention, induction of spontaneous linguistic behaviours and decrease of stereotyped and repetitive behaviours, although there are still limitations regarding the assumptions made due to the little number subjects tested in each study.
Following more of an engineering perspective, Cabibihan et al. [29] provided a brief technical datasheet of all social robots used in ASD interventions until 2013. The work addresses three key points concerning robot design features that need to be optimized in order to meet the needs of individuals with ASD (like physical appearance, functionality, level of autonomy among others) and associated effects, the set of behaviours to be stimulated in the child during the intervention (eye contact, imitation, joint attention, turn-taking, emotion recognition and expression, self-initiated interaction and triadic interactions), and discusses the roles and therapeutic benefits of social robots for children with ASD (as a diagnostic agent, friendly playmate, behaviour eliciting agent, social mediator, social actor or personal therapist). The authors’ results for the review’s key points were obtained through inputs from the actual user-end group (children with ASD) as well as with feedback from experts, therapists and parents in order to ensure the sustainability of a robot in the context of ASD therapy.

Table 3.2 is presented to summarize the relevant robot design results from Cabibihan et al. [29] review.

The revision presented in [33] by X. Liu et al. in 2017 also adopts an engineering perspective but focuses in a "systems and program" analysis for delivering ASD therapy, instead of robot design and target behaviours. As a matter of fact, this revision views technology used in autism studies as a human-machine interactive system model. In this review, the authors analyzed a series of studies under 4 key-points: (1) Presentation of the programs to participants with ASD, whether through computers, game consoles and mobile devices, virtual reality systems or social robots; (2) Input/reactions from participants, using a keyboard, mouse, touch screen, facial expression tracking, eye gaze tracking, speech recognition, human motion tracking through depth sensors among others; (3) Program customizability and adaptability, through automatic adaptation of programs or manual control and finally (4) Program evaluation, by behavioural pattern assessment, evaluation of treatment effectiveness and usability assessment.

Indeed, this revision gives a good insight of the state-of-the-art technology and programs used for ASD therapies. After analyzing the research results from more than 100 studies in the last five years (2012-2017) according to their key-points, the authors report that the only
Table 3.2: Robot design features [29]

<table>
<thead>
<tr>
<th>Robot Design Features</th>
<th>Function or Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appearance</td>
<td>Humanoid robots facilitate the generalization of the skills taught in therapy to real situations. More exaggerated facial expressions can enhance participant engagement and comprehension.</td>
</tr>
<tr>
<td>Sensory rewards</td>
<td>Providing explicit positive feedback after the correct execution of a task, for example through music playing, lighting up of the robot’s body part, or robot clapping is highly beneficial for keeping the child motivated and intrigued.</td>
</tr>
<tr>
<td>Locomotion</td>
<td>Mobile robots attract the child’s attention, allowing for enhanced playing scenarios.</td>
</tr>
<tr>
<td>Choice and control</td>
<td>Giving the child the ability to make choices during the interaction with the robot can keep him/her more interested in “making the interaction happen”.</td>
</tr>
<tr>
<td>Safety requirements</td>
<td>Robot should have no sharp edges, no rough movements and a minimum probability of malfunctioning, in addition to a robust design to endure being dropped on the floor or thrown to a wall.</td>
</tr>
<tr>
<td>Autonomy</td>
<td>The need for control every action of the robot should be avoided, and replaced for high-level control of the robot’s behaviours. Complete autonomy is not desirable.</td>
</tr>
<tr>
<td>Modularity and adaptability</td>
<td>The robot should have different functionalities and features to sustain each child’s different interests.</td>
</tr>
</tbody>
</table>

technology mature enough to capture the participant’s responses is eye gaze and that in order to assess how well a participant performed in the therapy session, the most common approach relied on video taping and offline analysis. Furthermore, the authors also pointed out that there are only a few autism treatment programs that are customizable and automatically adaptive whereas the large majority relies on Wizard of Oz control, which does not constitute a scalable solution and can affect the repeatability of the experiments. Finally, although children with ASD are generally more receptive to technology and show less anxiety when interacting with social robots, it is unclear if there are long-term benefits in gaining social and emotional skills, due to the lack of longitudinal studies on the long-term effectiveness of robot-technology for ASD therapies.
3.2.1 Important remarks

At the current state of the art, it seems robotic therapy has brought so far positive results, whether the robots are used as mediators, attractors or simply as measurement tools. Research has shown that using robots as therapy tools for ASD seems to improve engagement and elicit new social behaviours from children and adolescents with ASD [31, 23, 5, 30] such as joint attention and imitation.

In what concerns robot design, highly realistic human-like robots have important roles to fill, particularly for generalization of skills learned in therapy, although the trend seems to be moving towards visually and kinetically simple designs, especially given the sensitivity of children with ASD and the possibility of sensory overload.

It is noticeable that the most common approach for using social robots in ASD treatments is the Wizard of Oz scheme [30, 31] which gives the illusion that the robot can capture and understand the participant’s input/response and react properly, while in reality the robot is being controlled by a therapist or experienced person and the participant’s input/responses are collected by the robotic system or the therapist. On one hand, a WoZ scheme is not viable for a widespread robot use since it requires a human to control the interaction and can compromise the repeatability of a certain stimulus when a scenario is detected, which is relevant for learning social skills. On the other hand, some argue that complete autonomy is not desirable [29] given that technologies for detecting the participant’s emotion and actions and responding accordingly are not sufficiently reliable [33]. Thus, a WoZ setup offers a safe option and it can be designed to respond to high-level commands from the therapist to drive the general direction of the interaction, instead of a step-by-step control.

Currently, there are still some limitations regarding the use of social robots for clinical ASD treatments. Begum et al. [34] presented a review of the research for using social robots for ASD treatments until 2016, in which the authors’ pointed out that there is little progress in making social robots clinically useful for ASD treatment as well as methodological guidelines for that purpose, which will be explored later on. Scassellatti et al. [31] and Liu et al. [33] also referred that there are no large-scale longitudinal studies with many participants that provide quantitative measures, which makes it difficult to mea-
sure the benefits of design decisions for ASD therapies and to analyze the long-term implications in such a protected group of individuals as children with ASD.

### 3.3 HRI therapy for developing JA in children with ASD

As referred before, one common impairment in children with ASD is joint attention: demonstrating shared interest towards objects and others by pointing, referencing or using eye contact. In this section, five HRI studies that focus in developing JA skills in ASD therapy are analyzed.

It is noticeable that most studies that aim to develop joint attention skills through interaction with a robot are based on a series of prompts to elicit joint attention towards some target, either via human administrator or robot administrator, sometimes followed by a positive reinforcement related to the requested action (usually a sensory reward), in order to engage and motivate the participant.

The majority of these studies were also conducted with a Wizard of Oz human control scheme [3, 35, 36] or other forms such as teleoperated control [37], thus giving the illusion of an apparent autonomous robotic system while actually the system is at least partially controlled by human administrator, except the for study [38], which relied in a simple finite-state automaton.

Table 3.3 is a compilation of relevant studies for developing JA skills in the context of ASD therapy with a short description of the participants, robot used, therapy model, measurements performed and relevant findings of each. Figure 3.1 illustrates the robots used across studies from table 3.3. The different therapy models that have been used in experiments include:

- Single/Repeated experiment: The experiment may be performed once or across a number of sessions.
- Structured/Free-form interaction: Structured interactions are guided by the therapist as an active member in the experiment, whereas free-form interactions allow the child to play freely without intervention from therapist, unless necessary.
(a) NAO robot  
(b) Tito robot  
(c) Keepon robot  
(d) KASPAR robot

Figure 3.1: Robots used in ASD therapy

- Individual/Group interactions: The experiment may be performed on an individual or in several individuals.

Overall, by analyzing the findings in table 3.3 it seems that the use of robots for assisting ASD therapy has brought positive results in developing new JA skills and awareness of others’ presence, besides the increased engagement noted in human-robot interaction compared to human-human interaction. The revision by Scasselatti et al. [31] also reinforces this observation, stating that children with ASD who previously displayed tendencies to avoid eye contact or engagement with unknown adults demonstrated spontaneous joint attention behaviours when interacting with robots, such as looking at an adult and back to the robot, or pointing to the robot and looking at an adult or another child, with intention of sharing some feature with that person.

As [38] refers, joint attention skills within ASD depend on the in-
Table 3.3: Compilation of relevant studies for developing JA skills

<table>
<thead>
<tr>
<th>Study</th>
<th>Participants</th>
<th>Robot</th>
<th>Therapy model</th>
<th>Measurement type</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anzalone et al. [38]</td>
<td>16 ASD, 14 TD children</td>
<td>NAO</td>
<td>Single experiment; Structured; Individual interaction; No positive reinforcements</td>
<td>3D motion tracking system</td>
<td>Children with ASD turn their attention to the robot but both groups show lower performances in JA tasks if experimenters use a robot instead of the therapist. TD children have better performances.</td>
</tr>
<tr>
<td>Duquette et al. [37]</td>
<td>4 ASD children aged 5 (low-functioning)</td>
<td>Tito</td>
<td>Repeated experiment; Structured; Individual interaction</td>
<td>Behavioral observation</td>
<td>Children turn their attention to the robot and improve their performances in JA tasks if experimenters use a robot rather than human therapist.</td>
</tr>
<tr>
<td>H. Kozima et al. [35]</td>
<td>25 TD children ages 0-2 year-old; Children with ASD ages 2-4 in day-care centre</td>
<td>Keepon</td>
<td>Repeated experiment (throughout 3 years); Free-form interaction; Individual and group interaction</td>
<td>Behavioral observation</td>
<td>Simple robots like Keepon, only capable of direct gaze or joint attention, facilitate children with ASD to spontaneously exchange and share mental states with others, meaning that children with ASD possess the motivation for this.</td>
</tr>
<tr>
<td>B. Robins et al. [36]</td>
<td>3 children with ASD (low-functioning)</td>
<td>KASPAR</td>
<td>Repeated experiment; Free-form interaction; Individual experiment</td>
<td>Behavioral observations</td>
<td>Children were able to generalize their behaviour with KASPAR to co-present others and demonstrated awareness of the other’s perceptions of KASPAR. The robot acted as an object of JA (one child gazed and smiled at the therapist in response to KASPAR).</td>
</tr>
<tr>
<td>Z. Warren et al. [3]</td>
<td>6 TD, 6 ASD children with ages 3-5</td>
<td>NAO</td>
<td>Repeated experiment; Structured; Individual interaction; Provided sensory reinforcements</td>
<td>Number of prompts, eye gaze tracking with Tobii, targets identified</td>
<td>Children with ASD demonstrated improved performance across sessions, responding to JA prompts for looking at targets and sustained interest in the robot. TD children had better performances in looking at the targets.</td>
</tr>
</tbody>
</table>
teraction partner and the type of JA induction method. Multimodal joint attention induction (combination of instructions, gestures, visual or sound cues) appears to be the most efficient way of requesting joint attention towards a target for both TD and ASD children. Besides multimodal JA requests, positive reinforcements after correctly gazing at a target (RJA) or at the end of the session also seem to have successful results in generating new responses. Furthermore, positive reinforcements can also be used to encourage children with ASD to initiate interactions and proactive behaviours.

In what concerns robot’s appearance, the study conducted in [37], highlighted that children with ASD improved JA skills when interacting with Tito robot, which suggested that Tito’s simple yet friendly, non-human or animal appearance had an engaging effect in the user-end group. Robin et al. study [36] notes that KASPAR’s human features (eyes, nose, mouth etc) facilitated the generalization of the participant’s behaviours to at least co-present others. The study presented in [35] also argued that very simple robots like Keepon, with minimal and comprehensive expressiveness, only capable of looking to the participant or a specific object and demonstrating excitement, facilitated spontaneous behaviours and sharing of emotional states with others from children with ASD.

Although children with ASD tend to have better performances in displaying JA and other behaviours when using robots in therapy, as [32] states, there is lack of observation of these learned skills outside therapy sessions. The appearance of the robot (humanoid or non-humanoid) as well as the form and content of the therapy, hold great potential for the success of the intervention and generalization of the behaviours.
Chapter 4

Joint Attention Task

The MagPuzzle task was developed in the context of one of the current research topics of the Babyrobot project \(^1\) in the Speaking, Music and Hearing (TMH) department at Kungliga Tekniska Högskolan (KTH), with the intent of studying the effects of joint attention, mutual gaze and visual perspective taking in human-robot interaction.

4.1 Task Description and Hardware

The MagPuzzle task is a spatial reasoning task that uses physical magnetic pieces of different colors, which have a satisfying click when attached to each other. The playful nature of these pieces and the its target age group makes this task more translatable to the user end group, TD and ASD children [39].

The goal of this task is to visualize a three-dimensional cube and to reconstruct it in two-dimensions, connecting the magnetic pieces on top of a board with a 4x4 grid, before lifting the structure to assemble the final cube. There are three difficulty levels within MagPuzzle regarding restrictions on the maximum number of pieces used per line (row and column) to complete the two-dimensional cube. Throughout the task, a real life-like robotic head interacts with the participants as a peer, guiding them through the task’s instructions, helping to find a solution and providing feedback about the pieces played.

The feedback and the hints provided by the robot can be explicit, \(^1\)http://babyrobot.eu
through verbal cues synchronized with non-verbal gaze behaviours, or non-explicit taking the form of simple gaze behaviours, which are particularly relevant to establish joint attention with the participant. It is worth highlighting that the robot’s guidance is more successful if the participant establishes joint attention with the robot but it is not crucial for completion of the task.

The hardware setting for this task was designed to be simple and non-intrusive, making it suitable also for younger age groups. Unlike many studies in this area, users were not requested to wear special glasses or dedicated microphones, which in some sense limits the performance of the multimodal perception system but, on the other hand, it does not affect the naturalness and flow of the interaction.

Indeed, the hardware setting for MagPuzzle consists of two separate setups, the Wizard-of-Oz setup and the main setup. The Wizard-of-Oz setup is composed by a dedicated computer and a screen equipped with a commercial eye-tracker from Tobii \(^2\) to control high-level elements of the interaction. The main setup is located in a separate room where participants complete the task interacting with the robot. A back-projected robot-head (Furhat robot) \(^40\) is placed behind the table that contains the puzzle board and the robot’s height is adjusted to make the interaction on the same eye-level as the participant’s, accommodating most heights. Our robot has two degrees of freedom for neck movements, pan and tilt, and is capable of digitally displaying quick eye gaze changes, which are essential for this type of task where the users establish joint attention with the robot and in order to understand the robot’s visual perspective.

In front of the robot, we are using an RGB-D camera SR300 Real Sense from Intel \(^3\) to track the participant’s head position and orientation, which is capable of discriminating the user’s gaze between four quadrants in the board (top-right, top-left, bottom-right and bottom-left) and towards the robot. For object detection, the setup relies on an RGB camera placed on top of the board that detects fiducial markers placed on the pieces. For speech recognition and audio processing for detecting the participant’s utterances, a high quality amplified microphone is placed on the side of the robot.

---

2https://www.tobii.com
3https://software.intel.com/en-us/realsense/sr300
The overall system was implemented as a component-based distributed architecture capable of interacting via network in real-time. For clarity, the following list presents an overview of the main setup system components, enumerated in figure 4.1.

1. Furhat robot
2. Puzzle board
3. RGB Top camera
4. RGB-D camera SR300 Real Sense
5. Amplified microphone
6. Puzzle Pieces

Figure 4.1: Front view of the main MagPuzzle setup
4.2 Dialogue Acts and Wizard Interface

The system’s dialogue acts are used to instruct the participant about the rules of the puzzle, to provide feedback on the pieces played, to offer suggestions of where to play the next piece, to compliment and motivate the user and to answer simple questions, all in English language. These dialogue acts can include one or several behaviours, which are composed either by a sequence of gazes or by verbal cues synchronized with gaze patterns. When a dialogue act is chosen, one of the corresponding behaviours is selected, based on the assigned probability, and is then parsed by a behaviour planner that outputs the multimodal commands at the right time to the robot. Table 4.1 contains illustrating examples of the robot’s dialogue acts, where ‘X’ represents the vocalization of the color of the target piece.

<table>
<thead>
<tr>
<th>Speech Act</th>
<th>Content</th>
<th>Gaze Target</th>
<th>Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliment</td>
<td>“I think you are definitely on the right path”</td>
<td>Player</td>
<td>Happy</td>
</tr>
<tr>
<td>Bad Color</td>
<td>“The ‘X’ piece here is not likely to help to find the solution”</td>
<td>Square</td>
<td>Concerned</td>
</tr>
<tr>
<td>Bad Last Move</td>
<td>“That move will probably not bring us closer to a solution”</td>
<td>Square</td>
<td>Sad</td>
</tr>
<tr>
<td>Hint Correct</td>
<td>“I think putting a piece here might help”</td>
<td>Square</td>
<td>-</td>
</tr>
<tr>
<td>Good</td>
<td>“Good move”</td>
<td>Player</td>
<td>Excited</td>
</tr>
<tr>
<td>Good Quadrant</td>
<td>“Focusing on this area makes sense”</td>
<td>Quadrant</td>
<td>-</td>
</tr>
<tr>
<td>Reassurance</td>
<td>“I am confident that we will find a solution very soon”</td>
<td>Player</td>
<td>-</td>
</tr>
<tr>
<td>Probe Help</td>
<td>“Would you like me to make a suggestion?”</td>
<td>Player</td>
<td>-</td>
</tr>
<tr>
<td>Rule Max3</td>
<td>“Remember the instructions, in this round we must have 3 pieces in a line, but not more”</td>
<td>Player</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4.1: Illustrating examples of dialogue acts behaviours
The set of dialogue acts and corresponding behaviours implemented for this task were carefully elaborated to portrait a friendly cooperative robot that often smiles, displays emotions through different facial expressions, can refer to the pieces by their colors and to the participants by their name.

The wizard interface, implemented using the platform Unity game engine from Unity Technologies\(^4\), displays the set of dialogue acts through buttons with its respective labels (Speech Act on table 4.1) which, when clicked by the wizard, appear as subtitles in the interface, because the wizard does not have access to live audio or video streams of the interaction. In fact, the interface is restricted to just what the robot can perceive, forcing the wizard to adopt a more accurate visual perspective of the robot and to decide how and when to act solely with that information.

Figure 4.2 illustrates the WoZ interface, where the dialogue act’s labeled buttons can be either displayed directly on the interface, for delivering instructions or answering simple questions, or appear as expandable elements associated with a piece or a quadrant on the board (to respectively provide feedback or give a hint), the state (in)correctness of the game icon, the user’s icon and robot’s icon.

Furthermore, in order to lower the wizard’s response times to new events on the game as much as possible and to make the interaction feel more natural, the Tobii eye-tracker automatically exposes the expandable options when the wizard is looking at a specific area and hides them if wizard gazes to a different location. Plus, to further facilitate the wizard’s actions, the mouse pointer automatically appears in the location where the wizard is looking.

Besides the Tobii eye-tracker, a helping algorithm was also implemented in order to reduce the wizard’s cognitive load. Indeed, the dialogue acts’ buttons displayed directly on the interface are hidden if no longer needed for the completion of the task, and more importantly, by continuously monitoring the state of the game, the interface displays the correct places for providing the hints, the state (in)correctness of the game and if any rule was broken.

Finally, the last design consideration for the interface was to place all of the multimodal elements close together, so that the wizard is

\(^4\)https://unity3d.com
intuitively aware of all the changes in the interaction. The users can communicate with the robot through speech, gaze direction and new piece moves, all of which are sensed and displayed on the interface in real-time.

![Wizard of Oz interface screenshot](image)

**Figure 4.2: Wizard of Oz interface screenshot**

### 4.3 Multimodal Perception

As previously mentioned in this chapter, *MagPuzzle* is a multimodal system, comprised by a speech recognition module and two vision modules, one dedicated to the participant’s attention and the other to the pieces played on the board. The results from these modules are processed in real-time and displayed on the wizard’s interface.

The results from the speech recognition are obtained from a commercial cloud service and displayed on the bottom left box of the wizard’s interface. While the participant’s utterances are being processed, intermediate results are displayed on the wizard’s interface followed by three suspension points. When the final recognition result is available, the punctuation is removed and the result remains on the wizard’s interface until a new speech recognition event takes place. Besides the visual feedback of the participant’s utterances, the results from speech recognition are also vocalized at two times speed in
text-to-speech. The addition of this audio feedback ensures that the wizard can still perceive what the user is saying even if focused on other modalities of the interaction.

For the vision module dedicated to tracking the participant’s focus of attention we are using GazeSense 5 software, which is a non-intrusive head direction tracker. The results from the participant’s head direction estimation are represented in the wizard’s interface as two semi-opaque lines from the participant’s icon towards the board or towards the robot’s icon, which represent the participant’s gaze in real-time. Given the limitations of the non-intrusive attention tracker, when the participant looks at the board it only distinguishes between four different points identified as quadrants (top-left, top-right, bottom-left and bottom-right) or the robot.

Using a single camera to precisely track the eye’s pupil proved not accurate for detecting gaze directed towards a horizontal surface, which led to the decision of only using head orientation data for the participant’s gaze estimation. Furthermore, research has shown that head orientation contribution in estimating overall gaze direction is around 70% [41] and relying solely on head pose to estimate human visual focus of attention while interacting with robots has been reported to lead to reasonable accuracy levels in real-world, real-time measures [26].

Finally, regarding the vision module for tracking the puzzle pieces played on the board, we integrated an open-source augmented reality toolkit that robustly recognizes fiducial markers on the pieces in real-time. When the system detects a new piece or loses track of a piece, the location is automatically updated in wizard’s interface to keep the wizard aware of the state of the game.

---

5https://eyeware.tech/gazesense/
4.4 Autonomous joint attention system

The robot’s attention system is automated to produce gaze shifts triggered by the selection of dialogue acts and real-time multimodal perception. Within any gaze shift of attention, the robot automatically moves its eyes towards the target and only afterwards the shift is followed by the movement of the neck. This feature ensures more natural and believable results and is default in the attention system.

Regarding the gaze shifts triggered by dialogue acts, it is worth highlighting that the onset of linguistic vocalizations are posterior to the robot’s eyes and neck movements towards the intended target. For instance, with dialog acts associated with a target on the board, whether a piece or a quadrant, the robot first shifts its attention to the target and only afterwards executes the behaviour. The robot can gaze precisely at each of the squares on the board, since it stands on a fixed position and the gaze shifts towards the different locations on the board were manually calibrated. Furthermore, if the dialog act is not associated with the board, then the robot turns the attention to the participant with information collected from GazeSense at 30 frames per second. However, if GazeSense looses the location of the participant, it is assumed that the participant is located in front of the robot in a central position. Finally, when the behaviour encoded in the dialog act finishes the robot performs an idle gaze shift.

Idle gaze shifts are generated by the system, randomly, with respective target and maximum duration between 1 and 3 seconds. The target of the random gaze shift can include the participant, one of the four quadrants on the board or a direction that makes the robot seem distracted or avoiding mutual gaze. The probability of choosing one of the targets is not uniform and was fine-tuned after extensive experimentation to achieve a natural and believable behaviour. In what concerns the maximum duration of the idle gaze shift, the decision considered the findings from a study which reported that the average time people engaged on referenced objects was approximately 1.9 seconds, with no difference if the referencer was a human or a virtual agent [42]. Furthermore, the decision for the maximum time also considered the transmitted feeling trying to avoid a short gaze duration, which can reflect shyness or avoidance, and long gaze duration, which
can seem menacing and uncomfortable, by reacting to most events in the interaction and by randomizing gaze behaviours after periods of inactivity.

When the robot is not involved in a dialog act, the multimodal perception events described in the previous section can also trigger new gaze shifts in the system. Nonetheless, these triggers have limited priority and the robot does not react to a trigger from a perception event if it is involved in a higher priority gaze shift. That is, unless an gaze shift event with higher priority takes place, or the time duration of the current gaze is over and an idle gaze shift is issued, the robot maintains its current focus of attention.

The system tries to actively respond to the multimodal perception events to establish joint attention with the participant. For instance, if a piece is placed, moved or removed from the board, the robot assumes that the piece is the participant’s current focus of attention and gazes at the piece location. When a speech recognition event is detected, the robot shifts its attention to the participant to simulate that it is paying attention.

A summary of the possible gaze shift trigger events of the system are presented in table 4.2 with respective priorities, where 1 is maximum priority and 5 minimum priority.

<table>
<thead>
<tr>
<th>Event</th>
<th>Actor</th>
<th>Priority</th>
<th>Target(s)</th>
<th>Time</th>
<th>Condition(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialog act</td>
<td>wizard</td>
<td>1</td>
<td>Player, puzzle piece, quadrant</td>
<td>Behaviour length</td>
<td>Control, full</td>
</tr>
<tr>
<td>Puzzle piece moved</td>
<td>Player</td>
<td>2</td>
<td>Puzzle piece</td>
<td>1-3 sec</td>
<td>Full</td>
</tr>
<tr>
<td>Speech activity</td>
<td>Player</td>
<td>3</td>
<td>Player</td>
<td>1-3 sec</td>
<td>Full</td>
</tr>
<tr>
<td>Attention shift</td>
<td>Player</td>
<td>4</td>
<td>Player, quadrant</td>
<td>1-3 sec</td>
<td>Full</td>
</tr>
<tr>
<td>Idle gaze shift</td>
<td>System</td>
<td>5</td>
<td>Player, quadrant, distracted</td>
<td>1-3 sec</td>
<td>Control, full</td>
</tr>
</tbody>
</table>

Table 4.2: List of gaze shift trigger events with respective priority and enabling condition
Chapter 5

Methodology

This chapter describes the methodology followed by two MagPuzzle studies, differing on the end user group: adults and children. The use of a similar methodology for the two studies with the same MagPuzzle task, presented in the previous chapter, establishes the opportunity for a valid comparison of the results obtained between different age-groups and strengthens the conclusions drawn from the comparison.

Both studies were carried out by a researcher from TMH department (KTH) together with the author of this thesis, who performed the wizard-of-oz role in the MagPuzzle task.

5.1 Goal of the studies

The goal of each of the two studies, with adults and with children, is to assess the overall success of the joint attention system and to test the effects of visual perspective taking on the participants. Therefore, qualitative and quantitative measures need to be evaluated, for assessing the impact on the participants.

Furthermore, investigating performance differences between the two different cognitive ability groups (adults and children) presents some insight about the system’s design decisions that need to be readjusted to establish a more effective engagement with other cognitive ability groups and its social characteristics, as, for instance, children with ASD.
5.2 Participants

The adult’s study gathered 29 subjects from KTH university, where the experiment took place, with ages ranging from 18 to 34 years-old with an average of 22 years-old. They were mainly bachelor, master and Ph.D. students, recruited via flyers placed on campus or social media posts and were rewarded with a cinema ticket for participating in the experiment. However, technical issues and a change in protocol after the initial sessions reduced our initial pool to 22 participants, 11 males and 11 females.

Regarding the children’s study, 39 children from Internationella Engelska Skolan in Nacka (Sweden), between fifth and seventh grade, participated in the experiment, with ages ranging between 10 and 12 years-old. Our team asked the school to gather children in this age range who were willing to cooperate and whose parents authorized the participation in this study, which took place within Internationella Engelska Skolan facilities. Due to technical issues, this study could only consider 32 participants, 17 males and 15 females.

5.3 Procedure

In what concerns the adult’s study, in the beginning of the experiment the participants entered a sound proof room where the main setup was located, separated from the wizard. They were asked to fill in a consent form and an initial questionnaire. Afterwards, the experimenter provided a brief explanation about the nature of the MagPuzzle task and instructed each of the participants to sit down in front of the robot and the board. At the beginning of each interaction, the robot explains in detail the task and each participant completes the three puzzles with different difficulty levels in a time span varying between 5 to 10 minutes, depending on the participant. Finally, after the interaction finished, they were asked to fill in a final questionnaire about the interaction, which concluded the experiment.

For the children’s study, a similar procedure was carried out. However, the children were not requested to fill in an initial questionnaire.
nor the consent form, which had already been filled out by each child’s guardian. The final questionnaire was a slightly reduced version of the one used in the adult’s experiment and the total duration of the interaction varied between 7 to 15 minutes, depending on the child.

5.4 Manipulation

For assessing the overall success of the joint attention system and the effects of visual perspective taking on the participants, a separate operation mode was developed restricting the responses to joint attention, the control mode. In this operation mode, the autonomous joint attention system ignores the participant’s joint attention directives for generating gaze shifts (table 4.2). The wizard still initiates joint attention by selecting dialog acts targeted towards the board or the player, but in this mode the wizard’s interface does not display the participant’s gaze direction. The idle behaviours in this mode remain the same, creating believable patterns of gaze which can evoke feelings of naturalness and make the robot seem less mechanical, although the robot’s behaviour may appear more detached from the participant in this mode.

Thus, for investigating the previous hypothesis and for assessing the overall success of the joint attention system as well as the effects of visual perspective taking, participants were randomly divided and tested according to the two following conditions without their knowledge (single-blind experiment).

- **Control**: the robot only initiates joint attention (IJA)

- **Full Joint Attention**: the robot initiates joint attention, responds and follows the participant’s attention (IJA+RJA)

5.5 Measures

In the adult’s MagPuzzle study, the initial questionnaire reported demographic data, how often the participants are exposed to different technologies (video games, robots, virtual assistants) and their personality through a reduced version of the big-five questionnaire [43]. The
children where not asked to fill in the initial questionnaire because the nature of the questions is not suitable for children who are still in the process of developing their personality.

For both studies, the final questionnaires filled in by the participants provided a qualitative evaluation of the joint attention system. This questionnaire contained several Likert scale items that measured task difficulty, enjoyment, helpfulness, alliance and social presence (1: not at all; 5: very much). For measuring difficulty, in the adult’s study, we averaged three items where the participant reported the difficulty faced in each round, whereas with children the difficulty was assessed based on a single general item. For measuring enjoyment, helpfulness and alliance we chose a set of 8 questions that were tested for reliability. For measuring social presence we recurred to the social presence measure Networked Minds described on chapter 3 [28], a validated 36-item questionnaire that measures the degree to which an individual feels interconnected with another identity. To achieve a feeling of interconnection, participants should believe that both the robot’s behavioral and psychological involvement is connected to theirs and that their behavioral and psychological involvement is connected to the robot. The bidirectional nature of this measure makes it a sound choice to evaluate the effectiveness of our manipulation.

Besides from the qualitative assessments from the questionnaires, we video recorded each interaction and logged all the data collected by the joint attention system, in order to obtain more objective results about the effectiveness of the joint attention system. For this purpose, the data contained information about all the gaze shifts of the robot and the gaze shifts of the participant, as well as the speech recognition results, the puzzle piece moved and the associated (in)correct state of the game, the dialogue acts selected by the wizard and its eye gaze tracking results on the interface.

For the objective measurement of dyadic joint attention between the robot and the participant we considered two situations: mutual gaze (robot and participant looking at each other) and shared attention (both looking at the same object). This definition is used across conditions to compare the average percentage of time the robot established joint attention with the participant, similarly to an averaged With-ness measure [26] described in chapter 3, which also aims to measure
the extent to which a human is with the robot.

Finally, to operationalize objective information for the effects of visual perspective taking on the participants, we measure the success rate of the hints and feedback provided by the robot, given that the ability to respond correctly to these requires that the participant adopts the robot’s viewpoint of the board.
Chapter 6

Results and Discussion

The quantitative and qualitative results reported in this chapter were obtained through processing of the logged data and the questionnaire results from each interaction.

Before starting the analysis of the results from each study, it is necessary to highlight the difference of the robot’s behaviour between the two manipulation conditions, control and full joint attention. For this purpose, we objectively evaluate the average robot gaze distribution between both conditions, and also perform a third-party observer study.

The average robot gaze distribution was firstly investigated with the logged data from the adult’s study which is illustrated in figure 6.1, with N=11 participants for each condition. A Student’s t-test indicated that the average robot’s gaze towards pieces was significantly greater in the full joint attention condition (M =.31, SD=.08) compared to the control condition (M=.16, SD=.09), t(20) = -4.02, p=0.001. On the other hand, the average robot’s gaze towards the player was significantly higher in the control condition (M=.71, SD=.09) compared to the full joint attention condition (M=.59, SD=.09), t(20)=3.08, p=0.006. As expected, these results highlight the difference in the robot’s gaze behaviour between both conditions and reflect the increased robot’s attention to the board (pieces and quadrants) in the full joint attention condition, which was intended in order to respond to the participant’s attention. Similarly significant results were also obtained from the children’s logged data for the same gaze targets, which reinforces the previous observation.
For the third party observer study we created two videos, both showing the same actor interacting with the robot. The videos were both approximately 1:50 minutes long and displayed the same behaviour from the human and robot sides, however in one of the videos the actor was interacting in the control condition and in the other in the full joint attention condition. Then, we created a crowd-sourcing task and asked 40 participants to report if they could perceive differences in the robot’s behaviour in the two videos. Specifically, we asked the participants 7 questions where three of them concerned differences between the two videos, whether the they could perceive a difference in how well the robot was collaborating with the human, how well the robot was following the progression of the task and how helpful was the robot towards the human actor, and the remaining four questions covered four dimensions of the social-presence measure, described in chapter 3 [28]. The Co-presence and Message Understanding dimensions were excluded because they represent a construct that is harder to evaluate from an external perspective. Each dimension of the social-presence questionnaire was represented by a single item question, to adapt to the context and limitations of a crowd-sourcing task. We also gave the participants the possibility to indicate if they could not perceive any difference.

The results obtained in this single-blinded study are illustrated in

Figure 6.1: Average rate of time of the robot’s gaze distribution with adults.
figure 6.2, which exposes the frequency of third-party observers’ responses about which of the videos portrayed a more successful interaction according to the seven item questionnaire. Chi-square tests indicated significant differences between control and full joint attention conditions for items EU ($\chi^2(2, N=40)=12.95, p <0.01$), BI ($\chi^2(2, N=40)=12.05, p <0.01$) and help ($\chi^2(2, N=40)=7.85, p <0.05$), which confirm the positive trend in the full joint attention condition when focusing on the robot’s behaviour.

Figure 6.2: Frequency of third party observers’ opinions.

\[\begin{array}{cccccc}
\text{Third-Party Observer Ratings} \\
\hline
 & \text{AA} & \text{EI} & \text{EU} & \text{BI} & \text{collab} & \text{task\_prog} & \text{help} \\
\hline
\text{control} & & & * & & & & \\
\text{no difference} & & & & & & & \\
\text{full} & & & & & & & \\
\hline
\end{array}\]

1AA: Attentional Allocation; EI: Emotional Interdependence; EU: Emotional Understanding; BI: Behavioral Interdependence; collab: better collaboration ability; task\_prog: more awareness of the task progression; help: more helpful
6.1 Adult’s performance

The results obtained for the MagPuzzle study with adults are also reported in [44].

6.1.1 Objective Results

Analyzing the logged data from each interaction, we observed that the participants could successfully take the robots visual perspective, because the hints and feedback provided by the robot were correctly responded at a high rate. Indeed, the hints provided by the robot regarding the next move for the puzzle solution are given by head pose direction and eye gaze combined with an indirect verbal reference (e.g. ”I think putting a piece here might help”) and, although this type of cue was difficult to perceive due to the low resolution of the board with pieces placed together, the success rate for these hints was 63% across both conditions. Nonetheless, hints that combined head and gaze behaviours with explicit verbal cues, such as color features of the pieces (e.g. ”Removing this blue piece will bring us closer to a solution”), achieved a success rate of 93%.

Finally, to investigate if our manipulation in fact increased joint attention between the robot and the participant, we employed a Student’s t-test on the results obtained for average mismatched and shared attention, illustrated in figure 6.3. In these pie charts, precise joint attention refers to when the user and robot look at the same quadrant and imprecise joint attention considers the entire board. In this analysis we considered N=5 participants from control condition and N=8 participants from full joint attention condition, and, even with the low number of participants considered (a consequence of problem in logging the data in the first sessions), we obtained borderline significant results for both precise and imprecise joint attention. For precise joint attention, a borderline significant result was obtained in the full joint attention condition (M=9.57, SD=8.11) compared to the control condition (M=2.05, SD=2.12), t(11)=2.00, p=0.071. The same pattern was observed for imprecise joint attention, where we also obtained borderline significant result in the full joint attention condition (M=12.94, SD=8.11) compared to the control condition (M=5.53, SD=4.66), t(11)=--
Hence, these results suggest that our manipulation was successful in the sense that our system demonstrated more joint attention in the full joint attention condition. Moreover, a lower percentage of attention mismatch would probably not feel as natural as the robot would look too mechanical and exaggerated in responsive behavior.

6.1.2 Subjective Results

The qualitative results involve the social presence measure, validated in [28], and our measure of enjoyment, helpfulness and alliance. For the latter, we performed Cronbach’s $\alpha$ tests to assess the scale reliability of these scales for this study. The Enjoyment scale consisted of 2 items ($\alpha = .941$), the Helpfulness consisted of 3 items ($\alpha = .915$) and the Alliance, or Cooperation, consisted of 3 items ($\alpha = .816$). A Student’s t-test was performed on these three scales, and no significant results were found between conditions for the measures of Enjoyment and Helpfulness. However, Enjoyment appears to be inversely correlated with exposure to technology ($r(22)=-.455, p = .034$) and Helpfulness negatively correlated with how difficult participants considered the task ($r(22)=-.489, p = .021$). For Alliance we found borderline significant results between full joint attention ($M=4.67, SD=0.39$) and control conditions ($M=4.18, SD=0.74$); $t(20)=-1.93, p=0.068$.

For assessing the differences between conditions within the Social
Presence measure we employed non-parametric Mann Whitney’s U tests, as suggested by [28]. Figure 6.4 illustrates the average scores obtained for our measure, composed by our three scales and global social presence scales. In this figure, ”SP User-Robot” refers to the total perceived social presence connection with the Robot and ”SP Robot-User” refers to the total perceived robot’s connection with the Participant. A Mann Whitney test indicated that the overall perceived robot’s connection with the participant (SP Robot-User) was significantly higher in the full joint attention condition (Mdn=3.6) when compared to the control condition (Mdn=3.1), \( U = 27.5, p = .030, r = .46 \). However, in the opposite direction (SP User-Robot) there were no significant results.

![Figure 6.4: Mean values of the final questionnaire measures with adults](image)

In order to understand the cause of these results, we analyzed in further detail each of the six dimensions of the social presence measure in the two directions (Robot-User and User-Robot), represented in figure 6.5. In this figure, (SP-CO) co-presence, refers the degree to which the observer believes s/he is not alone; (SP-AA) attentional allocation, the amount of attention the user allocates to and receives from an in-
teractant; (SP-MU) perceived message understanding, the ability of the user to understand the message from the interactant; (SP-EU) perceived emotional understanding, the user’s ability to understand the interactant’s emotional and attitudinal states; (SP-EI) perceived emotional interdependence, the extent to which the user’s emotional and attitudinal state affects and is affected by the interactant’s emotional and attitudinal states; and (SP-BI) perceived behavioral interdependence, the extent to which the user’s behaviour affects and is affected by the interactant’s behaviour.

Indeed, Message Understanding ($U = 29.0$, $p = .037$, $r = .45$) and Emotional Understanding ($U = 33.0$, $p = .068$, $r = .39$) were greater in the full joint attention condition, being the most relevant dimensions for this effect. The most revealing item from our questionnaire that clearly shows and describes this effect is one of the items from message understanding, “My thoughts were clear to the robot” ($U = 12.0$, $p = .001$, $r = .70$).

Figure 6.5: Mean values of the six dimensions of the Social Presence measure [28] in each direction.
6.2 Children’s Performance

6.2.1 Objective Results

Similarly to the analysis with the adult’s study, we observed that the children could successfully take the robot’s visual perspective. Indeed, the success rate of the hints provided by the robot was 72% and the feedback cues that combined explicit verbal cues, such as the color of pieces, with non verbal behaviours achieved a success rate of 94%.

Moreover, in the full joint attention condition, this study also provided another type of hint liked to the participant’s gaze, the Quadrant hints. These Quadrant hints were given if the user’s gaze displayed on the wizard interface was focusing on the correct quadrant for the next move (e.g. "I believe you are focusing on the right place"), hence they could only be used in the full joint attention condition. The name Quadrant hint is related to the fact that the system can only discriminate the user’s gaze on the board between the four quadrants, as described on chapter 4. These hints achieved a success rate of 64% (in the full joint attention condition), which demonstrates the system’s ability to take the participant’s visual perspective as well, with a high success rate. Due to the a flaw in the wizard’s protocol, this type of hints was not used on the adult’s study for the full joint attention condition.

To assess if our manipulation also increased the joint attention established between the children and robot, we employed a similar approach as in the adults’ study. The results obtained for averaged the mismatched and shared attention are presented in figure 6.6, with N=15 participants in control condition and N=17 participants in full joint attention condition. A Student’s t-test indicated that the average time of precise joint attention between the participant and the robot was significantly greater in the full joint attention condition (M=15.91, SD=9.58) when compared to the control condition (M=4.66, SD=6.69), t(30)=-3.80, p=0.001. On the other hand, for the average time of mismatched attention between user and robot, a significantly greater result was found in the control condition (M=58.07, SD=28.39) when comparing to the full joint attention condition (M=32.95, SD=19.61), t(30)=2.94, p=0.006.

These results also highlight the suggestion that our system can suc-
cessfully establish more joint attention with the participant in the full joint attention condition, as well as decrease the attention mismatch in the interaction, which still maintains a believable percentage of time, avoiding too mechanic and less natural behaviours.

### 6.2.2 Subjective Results

The qualitative results analyzed in the children’s study, similarly to adult’s study, involve the validated social presence measure \[28\], our measure of Enjoyment, Helpfulness and Alliance or Cooperation, as well as an analysis of the visual perspective taking results with the social presence measure.

For our measure of Enjoyment, Helpfulness and Alliance no significant results were found between conditions. Furthermore, when employing non-parametric Mann Whitney’s \(U\) tests to assess the differences between conditions in the social presence measure, a trend for higher results in the full joint attention condition was observed, but no significant results were associated in any of the dimensions or the global scale. Figure 6.7 summarizes the observations made.

Although no significant results were found between conditions in the questionnaires, interesting correlations were discovered between the Quadrant Hints success rate and the Social Presence questionnaire. Indeed, these hints, which are only used in the full joint attention condition, are strongly positively correlated with the dimensions Co-
presence (0.62), Attentional Allocation (0.68), Message Understanding (0.68), as well as with specific directions of the dimensions, such as Co-presence Robot-User (0.73), Attentional Allocation User-Robot (0.70) and Message Understanding Robot-User (0.70) and more importantly, with the global scale Social Presence (0.76) and each of its directions, Social Presence User-Robot (0.69) and Social Presence Robot-User (0.80). These significant correlations can explain the slightly higher Social Presence results in the full joint attention condition, given that these hints are only used for this condition and also due to their success rate of responses.
6.3 Comparison and Discussion

The previous sections contrasted the results between control and full joint attention conditions to assess the effectiveness of our system. In this section, the main goal is to compare and analyze the performance differences in the *MagPuzzle* task between the two different age groups (adults and TD children), to have some insight on the robot’s behaviours and design decisions that are the most successful for younger age-groups. For this reason, all comparisons in this section refer solely to the results obtained in the Full Joint Attention condition, since it is the condition that includes all of the system’s functionalities and design decisions as intended to establish joint attention, besides that the baseline performance differences (control condition) fall out of this thesis’ scope.

6.3.1 Objective Results

As expected, children faced an increased difficulty in completing this task, when compared to adults. This is reflected by a number of factors, described in further detail on the following paragraphs.

Children took more time to complete the *MagPuzzle* task. Analyzing the average total playing time, a Student’s t-test with N=11 adults and N=17 children indicated borderline significant difference between children (M=418.5 sec., SD=185.5) and adults (M=313.9 sec, SD=79.7); t(23.39) = -2.05, p=0.052.

The increased difficulty can also be linked to the fact that children of that age are still developing spatial reasoning and, therefore, need more guidance from the robot and more feedback on wrong moves in order to assemble the cubes. A Student’s t-test with the same number of subjects confirmed that the number of incorrect moves across the three game rounds was significantly higher with children (M=53.18, SD=41.48) when compared to adults (M=30.09, SD=14.84); t(21.60)=-2.10, p=0.048. The same pattern was observed in what concerns the total number of hints given through the task. Children (M=5.65, SD=5.00) needed two times more hints than adults (M=2.45, SD=2.70), which a Student’s t-test also indicated as borderline significant with the same number of subjects; t(25.41)=-2.19, p=0.038.
Nonetheless, it is worth highlighting that the success rate of piece Hints had no significant difference between children (72%) and adults (63%). The same pattern was observed for the feedback hints related to piece color features, with no statistical significance (adults: 89%, children: 94%). All these results suggest that, although children have more difficulty in completing the task, both children and adults can successfully take robot’s visual perspective and correctly follow its guidance.

Analyzing the system’s gaze behaviour between age groups, a Student’s t-test revealed borderline significant differences in the robot’s gaze rate towards the player, which was higher in adults (M=0.59, SD=0.90) than in children (M=0.50, SD=0.93); t(26)=2.66, p=0.013, as well as the robot’s gaze rate towards pieces in the board, which was higher for children (M=0.36, SD=0.06) when compared to adults (M=0.31, SD=0.08); t(26)=-2.17, p=0.039. This effect can be explained by the fact that children needed a considerable amount of hints and feedback for completing the task when compared to adults, which required that the robot directed its gaze towards the board, whereas with adults the robot had more opportunities to look towards the player.

Finally, in what concerns the mismatched and shared attention differences between adults and children, the results are illustrated in figure 6.8. The statistical tests evaluated for these results considered N=8 adults and N=17 children, and, even with the low number of adult subjects in the full joint attention condition, due to technical errors and a change in protocol explained in section 6.1, some significant results were found between the two age groups.

As expected, a Student’s t-test indicated that there was a borderline significant difference in the established JA towards the Board (robot and participant looking at any place on the board) with a greater percentage for children (M=38.50, SD=20.57) than with adults (M=22.51, SD=15.44); t(18.06)=-2.16, p=0.044, which is a consequence of children needing more feedback and hints to complete the task and is in line with the increased robot’s gaze rate towards pieces. The same pattern was observed for Imprecise JA (robot and user looking at different quadrants on the board), with borderline significance for children (M=22.59, SD=12.51) when compared to adults (M=12.94, SD=8.12); t(20.30)=-2.31, p=0.032.

Although Mutual Gaze did not have any statistical significance, it
is visible that adults have a higher percentage of mutual gaze with the robot over children, as an indirect consequence of not needing so many hints. It is also noticeable that Mismatched attention, besides having believable percentages for each of the groups, conveyed quite similar results for both groups, which demonstrates the coherence in our system’s natural and human-like behaviour.

Overall, it seems that children had a better performance in establishing joint attention with the robot, given that during each interaction there were more opportunities for this than with adults, which suggests that the MagPuzzle task is more suitable for children in this age range, as it was originally intended [39].

### 6.3.2 Subjective Results

In order to investigate the qualitative differences between adults’ and children’s performances, we employed non parametric Mann-Whitney’s U tests for the social presence measure, as suggested by [28], and Student’s t-tests for our Enjoyment, Helpfulness and Alliance or Coop-
eration measure. All the analysis in this subsection consider the responses to the final questionnaires from \(N=11\) adult participants and \(N=17\) children participants (\(N=28\) total participants from the full joint attention condition).

The results of the final questionnaire are illustrated in figure 6.9 for both adults and children and it is noticeable that, overall, children attributed higher scores than adults in most of the scales, with exception of the Cooperation scale, which confirms that children tend to answer surveys in order to please the experimenter rather than truthfully express their opinion, as reported in [45].

Nonetheless, children were more clearly more enthusiastic about interacting with the robot, as expected, given that they did not had much contact with this type of technology compared to adults, and felt that the robot provided useful hints for completing the task. Indeed, a Student’s t-test indicated borderline significant results for the Helpfulness scale in children (M=4.63, SD=0.41) when compared to adults (M=4.06, SD=0.71); \(t(26)=-2.69, p=0.012\), as well as for the Enjoyment scale in children (M=4.88, SD=0.22) as opposed to adults (M=4.45, SD=0.65), \(t(26)=-2.52, p=0.018\).

For the Social Presence scales, a Mann Whitney’s U test revealed borderline significant results for the total perceived participant’s connection with the robot (SP User-Robot) with children (Mdn=4.0) over adults (Mdn=3.6) \(U=50.5, p=0.042, r=.38\), and for the total perceived robot’s connection with the participant (SP Robot-User) for children as well (Mdn=4.1) when compared to adults (Mdn=3.6) \(U=55.0, p=0.073, r=.34\).

Analyzing each of the six dimensions of the Social Presence measure, as well as the global scale, a Mann Whitney’s U test also revealed significant results in some of the dimensions, all of which follow the trend of higher results for children, even in the ones where the differences are not significant, as portrayed in figure 6.10.

As a matter of fact, the Message Understanding sub-scale, which reflects the ability of the user to understand the message from the interactant, was found to have a significantly greater result with children (Mdn=4.2) than with adults (Mdn=4.5) \(U=36.0, p=0.006, r=.52\). A Mann Whitney’s U test also indicated a borderline significant result for perceived emotional understanding (Affective Understanding) for
Figure 6.9: Mean values of the final questionnaire measures for adults and children in full joint attention condition

children (Mdn=3.5) over adults (Mdn=2.7) $U=49.5$, $p=0.037$, $r=.39$. The same pattern was observed for the Emotional Interdependence sub-scale for children (Mdn=3.3) compared to adults (Mdn=2.8) $U=42.0$, $p=0.015$, $r=.46$. The results obtained in these sub-scales are responsible for the borderline significant result on the Social Presence measure (total perceived social connection with the interactant) which, once again, was higher for children (Mdn=4.1) when compared to adults (Mdn=3.6) $U=42.0$, $p=0.015$, $r=.46$.

Overall, the results obtained for both groups suggest that the MagPuzzle task is more adequate for children in the age-range of 10 to 12 years-old than adults, given that our system could successfully establish more joint attention with this group, and also given the higher enjoyment, helpfulness and social presence scores that children reported after the interaction with the robot. It was also observed that children of this age could understand the robot’s visual perspective of the board and react with the same success rate as adults.
Figure 6.10: Mean values of the six dimensions of Social Presence measure for adults and children in full joint attention condition
Chapter 7

Guidelines for robotic systems to assist children with ASD

Given the impairment that children with ASD face in establishing joint attention and taking someone else’s visual perspective, this chapter aims to present some design guidelines for robotic systems, similar to MagPuzzle, to assist in joint attention therapies for children with ASD, which are elaborated from the results obtained in the previous chapter for the performance differences, literature research, together with the input from an interview with an autism therapist.

It is worth to highlight that the role of the robotic system is not to substitute the therapist but to assist in the treatment throughout the therapy session, as children with this pathology demonstrate higher motivation and proactive behaviours when interacting with robots [5, 23, 30] as exposed in chapter 3.

There has been a growing interest in designing interactive robots that human children can naturally interact with. Indeed, socially assistive robots for ASD therapies aim to generate one or more carefully designed interactions between human users and themselves, involving elicitation, coaching, and reinforcement of social behaviour.

In order to have a good insight about robot-mediated interventions for children with ASD, it is useful to have some knowledge about general game/activity design features appropriate for this purpose. For this matter, in [46] Bartoli et al. presented some game design guidelines, gathering knowledge from several specialists including neuro-
logical doctors, therapists, educators and their team. These guidelines were presented in the context of motion-based touchless games but the general rules are applicable for other sorts of games, and they include the following key-points:

- **Strong customizability.** The game must be customizable to fit the needs and preferences of each child with ASD, given the different strengths and skills deficiencies characteristic of this pathology.

- **Increasing levels of complexity of game tasks.** As a child acquires more skills, progressively more challenging tasks should be made available.

- **Clear and easy to understand task goals.**

- **Multiple means of communicating game instructions, such as text, voice and visual cues.** Some children with low functioning ASD typically need visual cues in particular.

- **Positive reinforcement with rewards.** Game score alone might not be enough to motivate children with ASD. Hence, other forms of reward such as video or audio effects should be provided to encourage and motivate participants.

- **Repeatability and predictability of game play.** Unpredictability may cause anxiety to children with ASD and repeatability is needed for the participants to learn.

- **Smooth transitions.** The transitions to a higher level of the game should happen without noticeable delay so that the children with ASD are not discouraged.

- **Minimalistic graphics and sound/music.** All graphical/sound elements must be included in the game goal due to the fact that children with ASD may be subject to sensor overload.

- **Dynamic stimuli.** Prolonged static scenes should be avoided to trigger motor rigidity.

Furthermore, one common characteristic of ASD is impairment in the use of eye-to-eye gaze, joint attention, facial expressions and other
social behaviours that regulate engagement. Correspondingly, a design goal adopted by many SAR systems is eliciting and maintaining active engagement throughout an interaction, like MagPuzzle system, for instance. For developing specific social skills in ASD therapy such as joint attention skills, [46] advises to include in a task one or more of the following situations:

- One child has to call the attention of another child, or adult, towards target objects, whether through pointing, eye gazing or vocalizing (IJA).

- Two children have to coordinate each other in order to find a target object and catch it simultaneously. (or one child coordinating to find the target with help from a robot: IJA+RJA)

- Two children (or child and robot) have to alternate in the same or different tasks (“turn-taking”) and regulate one’s behaviour to the one of the co-player.

### 7.1 Interview with Autism Expert

The author of this thesis had the opportunity to interview an ABA (Applied Behaviour Analysis) autism therapist expert, Dr. Catharina Hamilton, who kindly accepted the invitation. This informal meeting started with the demonstration of a video with a MagPuzzle interaction to the therapist followed by some feedback about the typical therapies used for training joint attention with children with ASD.

After the visualization of the MagPuzzle video interaction, the author asked the therapist about features or characteristics this setup and task should meet in order to correspond to the needs of an child with ASD for training JA skills. The therapist’s response focused essentially on the need for a positive, fun reinforcement after the correct completion of a request, in order to keep the child motivated. The reinforcement can be of any kind as long as it corresponds to the child’s previously known interests, for example a smile, an encouraging prompt, or even giving the child one of his/hers favourite candies, or the demonstration of a short video clip from one of the child’s preferred cartoons, among others. Furthermore, the therapist added that the task/activity
should ensure that child always succeeds, and even when the child
does not respond correctly to the stimulus, the reinforcement should
be provided anyway in order to avoid frustration.

In what concerns the feedback about the typical therapies used for
training joint attention, the therapist highlighted the need for each
therapy session to be adapted to the child’s interests and skills - in-
dividual and customizable therapies. Specifically, for training joint at-
tention skills, the therapist referred that the child should have baseline
eye-contact and pointing skills, and an IQ above 70.

Another relevant issue is the importance of therapy structure. Typ-
cical therapy sessions involve the presence of a therapist or a parent
together with a therapist so that the child feels in a comfortable, famil-
 iar environment. Moreover, before the therapy session starts, the child
should be given some free time to explore and get familiar with the
setup. The session itself usually follows the so called ABR-structure,
with illustrating examples presented on table 7.1.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother points towards bus</td>
<td>Child looks at bus and looks back at mother</td>
<td>Smile; Verbal encouragement + Enhancer</td>
</tr>
<tr>
<td>Parent/Therapist places object of interest in unreachable shelf</td>
<td>Child looks/calls for therapist or parent and asks to be given the object</td>
<td>Smile; Verbal encouragement + Receives object</td>
</tr>
</tbody>
</table>

Table 7.1: Illustrating examples of ABR therapy for training joint attention

Firstly, the parent or therapist provides a prompt eliciting a re-
sponse (A) which requires establishing joint attention. Then, the in-
tended behaviour from the child (B) has to include the response to the
joint attention bid plus looking back at the parent or therapist. How-
ever, if the child fails to respond properly to the elicitation, then an
extra prompt is provided between A and B, which can be a verbal
rephrasing of the elicitation prompt together with pointing, or even
some soft assistance in turning the child’s head towards the elicitation
target. Finally, the positive reinforcement (R) should be provided after
the correct response from the child, or after the second prompt, in case
of failure, to maintain the child motivated and collaborative, through verbal positive encouragement together with an enhancer that meets the child’s interests.

7.2 Extending Design Guidelines

This section presents an extension for the design guidelines laid out by [34] and the feedback provided the autism therapist.

Focusing on the design features that generated the positive results in the *MagPuzzle* studies, particularly with TD children, the author elaborated some general design guidelines hoping to provide some ground for future research in human-robot interactions with the therapeutic application of developing joint attention skills in children with ASD.

- Task with low level of verbal skills, that promotes joint attention behaviours. The task performed should be adapted to the end-user’s abilities. If joint attention skills are the target skill for the therapy and since these skills are prior to the development of speech, the completion of the task should not depend on speech input from the participant, to avoid frustration and to keep the child collaborative.

- Non invasive setup. Special eye-tracking glasses or dedicated microphones should be avoided in order not to intimidate children with ASD, who are particularly sensitive to new environments.

- Wizard-of-Oz control. A high-level control scheme is preferred over full automation with this user-end group. The state-of-the-art technology for detecting the participant’s emotions and actions and responding accordingly is not sufficiently reliable, and therefore a WoZ setup offers a safe option.

- Multimodal Perception. Similarly to a human-human interaction, the robot should sense speech, gaze and even the emotional state of the child, for having an increased social presence and for maintaining believable long-term interactions.
• Multimodal Behaviour Generation. The behavioral output of the robot should involve both verbal and visual cues for a higher engagement with the child and to facilitate the communication of the task goals. Children with low functioning ASD need visual cues in particular.

• Responding to Joint Attention. The positive results obtained in the MagPuzzle studies in the full joint attention condition suggest that a robot that reacts to the participant’s actions and inputs is more successful in establishing joint attention.

• Physical embodiment and Human-like robot appearance. TD children reported a higher level of enjoyment and helpfulness in the MagPuzzle studies when interacting with the robot, which suggests that Furhat’s physical appearance and proactive behaviours may have an increased positive impact in younger age-groups. Furthermore, a human-like appearance has been reported to be useful for generalization of the skills learned in therapy.

These guidelines focused essentially on robotic-system features for successfully establishing joint attention with children with ASD, which need future validation with this user-end group. Together with the need for a positive reinforcement and therapy structure highlighted by the autism therapist, the adaptation of the MagPuzzle as a therapeutic tool for the cognitive needs of children with ASD seems possible, given the positive results achieved with adults and with TD children.
Chapter 8

Conclusion

The research question covered by this dissertation was: How to evaluate and design autonomous joint attention mechanisms for different cognitive ability groups. Particularly, this research focused on adults, TD children and children with ASD, with their respective characteristics and cognitive abilities.

To answer this research question we started by a literature review of the design decisions in state-of-the-art human-robot interaction systems, with a special focus on the field of social robots that assist in ASD therapies. The review also covered an analysis of robot-assisted ASD therapies focused in developing joint attention skills for children with ASD, with the respective elicitation mechanisms for this matter, and finally, it presented a compilation of relevant joint attention measures, both objective and subjective which would be employed later on.

After the literature review, this thesis presented a detailed description of the design decisions of our semi-autonomous system that simulated joint attention in human-robot interaction, the MagPuzzle system. Then, in order to address the evaluation part of the research question we tested the effectiveness of our system across two different groups, adults and TD children. The effectiveness of our system is understood as successfully establishing joint attention with the users and the users correctly understanding the robot’s visual perspective.

The results obtained for the adult’s study were published in a paper for ICMI 2018: The 20th ACM International Conference on Multimedia Interaction [44], for which the author of this thesis collaborated as one of the authors, and is included as an appendix on this thesis.
Overall, the results obtained in each of the two studies with adults and TD children demonstrated that our system could successfully establish joint attention with the participants, which was evaluated through an objective joint attention measure, and that the participants could successfully perceive the robot’s view point, which was assessed by the success rate of the eye-gaze hints provided by the robot.

These results reflected an increased social presence felt by the participants and showed the importance of enabling robots with joint attention mechanisms that are not necessarily relevant for the completion of cooperative tasks. We believe these mechanisms will contribute for establishing and maintaining believable long-term interactions with robots that cooperate and sense the multimodality of the social world around them.

Furthermore, by comparing the results obtained between these groups, a higher level of enjoyment and joint attention were reported with TD children, which lead the author to the conclusion that the MagPuzzle system is in fact more adequate for children.

Finally, in order to probe the effectiveness of our system with children with ASD a different approach had to be followed, due to the necessary redesigning of the task, the sensitiveness of this user-end group and the ethical concerns this type of experiment could raise. The author of this thesis had the opportunity to interview an autism therapist, who after viewing a demonstration of an interaction with the MagPuzzle system, provided some feedback for adaptation features and also shared some of the typical games performed in therapy for teaching joint attention.

Through the results obtained from the MagPuzzle studies, the feedback from the therapist and the literature study, the author elaborated general system design guidelines for this purpose and feels that the application of these guidelines to the MagPuzzle task could be of great interest for a potential therapeutic tool in joint attention focused therapies.

Overall, this thesis demonstrated the effectiveness of our semi autonomous joint attention system, MagPuzzle, across two different user-end groups together with the perceived effects from the user’s side, and gave the first step towards the adaptation of this system for therapeutic applications focused in developing joint attention for children.
8.1 Final Discussion

This section intends to elaborate on the most adequate way to measure and evaluate Joint attention robotic systems according to cognitive group.

After analyzing the results of the studies for both adults and children between 10 and 12 years-old, it is clear that assessing the effectiveness of human-robot interaction with children is more problematic than with adults. While adults can be probed using questionnaires and self-reflections, children have the tendency to please the experimenter rather than truthfully answer the survey questions, which was also observed in [45].

In the MagPuzzle study with children the Likert scales came back with extreme results, which makes us believe that children tried to second guess the desired response and sometimes answered randomly. For this reason, alternative measuring ways need to be evaluated with TD children as well as children with ASD, which need to cover the goal of the interaction in order to assess its effectiveness. Global quantitative measures such as the interaction duration, participant utterances or compliance with robot suggestions can only be trusted so far, depending on the goal of the interaction.

For assessing the effectiveness of an interaction in establishing joint attention with children, as in the case of MagPuzzle, the most reliable measure probably falls on objective measures of time of established joint attention for each target.

However, if the goal of the interaction is educational, [45] advises balanced pre-interaction and post-interaction tests in order to assess what is contribution of the robot to gain knowledge. Therefore, for training joint attention skills with children with ASD through human-robot interaction, a measure of JA should be taken into account and followed by carefully designed pre and post tests, in order to assess the extent to which the interaction truly improved the child’s abilities - generalization of the skill.
8.2 Limitations and Future Work

In what concerns the development of a system for assisting in ASD therapy, with the goal of developing joint attention skills, this thesis has only laid out design guidelines, some of which are already followed by MagPuzzle.

The author is aware that the design guidelines exposed in chapter 7 for robotic systems to assist in ASD therapy focused in developing joint attention skills need to be validated with a study with the intended user-end group: children with ASD. Furthermore, for robotic-system interventions to be clinically useful for treating ASD, they have first to be considered evidence based practices (EBP) [34]. This concept comes from the medical field to minimize the gap between research and practice, allowing clinicians to choose methods with strong scientific evidences from carefully controlled research studies.

Therefore, for this purpose, the validation of the design guidelines studies on robot-mediated interventions for treating ASD should follow a systematic methodology, composed by six elements exposed in [34] by Begum et al:

1. Goal of intervention. What is the specific clinical goal the study or research seeks to achieve and why?

2. Participants. What type of participants does the research pretend to target?

3. Independent variables. The robot itself and its behaviours are independent variables. The hardware and software components should be described with enough detail and precision for replication.

4. Dependent variables. What specific ASD behaviours will be affected/changed by the intervention?

5. Research design. What is the most adequate research design to evaluate the goal/outcome of the intervention, given the number and type of participants?
6. Generalization training. What is the plan to help the participants generalize the skills learned in the intervention with the robot to other humans?

Following this line of thought, it would be of great interest if the *MagPuzzle* task could be adapted to the needs and the cognitive abilities of children with ASD, in order to assess the effectiveness of the interaction in improving the child’s joint attention skills.
Bibliography


Appendix A

Publication

A.1 ICMI 2018: The 20th ACM International Conference on Multimodal Interaction
Using Multimodal Perception to Improve Joint Attention in Human-Robot Interaction

Andre Pereira
KTH Royal Institute of Technology
Stockholm, Sweden

Catharine Oertel
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

Leonor Fermoselle
KTH Royal Institute of Technology
Stockholm, Sweden

Joe Mendelson
Furhat Robotics
Stockholm, Sweden

Joakim Gustafson
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT
Joint attention has been shown to be not only important for human-human interaction but also for human-robot interaction. It can help to make cooperation more efficient and support disambiguation in instances of uncertainty. In this paper, we present a semi-autonomous system that uses information from multimodal sensing to simulate joint attention in human-robot interaction. We investigate the effects of such a system on people's perception of a humanoid robot within a problem solving task. Results showed that the robot successfully helped and formed an alliance with human participants. In addition, by studying human perceptions of social presence towards the robot from both a first and third party perspective, we show the importance of enabling robots with subtle joint attention mechanisms that might at first glance appear as not necessarily relevant for the completion of cooperative tasks.

CCS CONCEPTS
- Human-centered computing → User studies; Natural language interfaces; Empirical studies in HCI; Empirical studies in collaborative and social computing;

ACM Reference Format:

1 INTRODUCTION
Humans are good at observing others and at understanding their mental states, such as what they are seeing, feeling and what they want. An essential social skill required to achieve this is to be able to spontaneously see and understand the world from someone else's viewpoint. This skill is called Visual Perspective Taking (VPT) [9], and plays a critical role in social cognition and interaction between humans. The same effect has been reported in Human-Robot Interaction (HRI) where people are similarly able to override their egocentric viewpoints and spontaneously take robots' visual perspectives [33]. If social interaction partners are aware that they are attending something else in common, that phenomenon is usually described as joint attention. Joint attention describes not only the visual perspective taking required for this phenomenon but also the tendency for social partners to focus on a common reference and to monitor the others' attention to an outside entity [30]. Identifying both verbally and non-verbally referred-to objects is extremely important for everyday collaborative tasks.

Body orientation, head, and eye motion are non-verbal cues or behaviors that carry a clear meaning in communication. These cues are relevant for establishing and maintaining a connection between two parties as they often reveal a participant's Visual Focus Of Attention (VFOA). Simulating attention shifts and having a believable VFOA model in social robots is fundamental for establishing and driving believable effective interactions between humans and robots [6]. Attention shifts occur when humans communicate, and situated dialog between humans and robots benefits from not only verbal but also these nonverbal behaviors. Nonverbal behaviors have been shown to increase task efficiency and improve collaboration between people and robots [2]. As an example, a robot in a factory assembly line should use both of these cues to reduce uncertainty and better guide the user into which object should be assembled next.

Assuming that people do take robots' visual perspectives, a perception of agency is also inferred from periods where no verbal conversation is present. In these moments, the robot still needs to shift its attention and display non-verbal behaviors to show that it is "alive" and listening. To succeed in this, it is essential for robots to not only simulate but also detect humans' attentional activity. In this paper, we describe a semi-autonomous system built to perceive and synthesize joint attention in a human-robot interaction tasks. Real-time autonomous mechanisms simulated in our robot are able to not only initiate joint attention, but also produce believable attentional behaviors that respond to the users' attention. Our system is evaluated within a spatial reasoning problem solving task. User perception studies revealed the positive effects of using multimodal perception to generate believable joint attention behaviors.

2 RELATED WORK
There are many studies which are concerned with multimodal behavior generation in virtual agents and robots [17, 24]. The focus
often lies on combining prosody, gaze and gestures in a human-like manner. Non-verbal behavior generation is not only important for creating the illusion of a human-like appearance and behavior but also for improving the task performance between humans and robots [2]. It has been found that a robot looking at an object while being instructed should at certain times switch its gaze to the instructor. A robot that seeks mutual gaze was perceived as featuring an improved stance when compared to one that randomly switches its gaze from the object under discussion to random places [14]. It achieves longer interactions, an increase in human attention and it gets more corrections from the instructor when it makes mistakes.

From the perception side, it has also been shown that using gaze-tracking to disambiguate spoken referring expressions improves the interaction with a robot by making it more natural, pleasant and efficient [19]. In human-human-computer interaction, it has been shown that continuously updating the partner’s gaze direction on a screen can help learners achieve a higher quality of collaboration and learning gains [26].

Collaborative robots also need to coordinate their gaze behavior with the human collaborator’s actions. This has been found to increase the human’s feeling of being looked at [32]. An example of a robot that coordinated its gaze with human actions was reported in [7]. In this experiment, humans taught a robot the names of three buttons that they later asked to push. It was found that a robot with gaze controlled by the human movements made it easier for subjects to understand the robot’s current state and abilities. This in turn led to better task efficiency and robustness to error.

One aspect which has proven to be a very important factor for task-based human-robot interaction is joint attention. Joint attention is an important factor for quite diverse reasons. It can better enable a human and robot to coordinate in an assembly task but also equally important when trying to teach children on the autism spectrum social and language skills. Indeed, there is a considerable amount of literature [15, 16, 21] that views joint attention as a pivotal skill, that is, a skill that when subject to intervention will result in positive collateral changes of other untreated skills, such as language and social-cognitive developments. Therefore, quite a number of studies have been carried out in this area. For example in [20], it has been shown that joint attention in the context of handing over a water bottle could improve the timing and perceived quality of the handover event. In [28] it has been shown that users are benefiting from joint attention with the robot when disambiguating landmarks in a route drawing task. It has also been studied how joint attention might be improved by using for example spatial references and action observation [22].

Similarly to our research, studies have indicated that human perceptions of robots that use joint attention can be positively evaluated. For example, [13] has found that robots that respond to joint attention are more transparent to humans and perceived as more competent and socially interactive. The study described in [31] found through subjective evaluations and observations of user’s gaze, that joint attention draws the user’s interest along with the user-guessed interest of the robot. The eye contact brings the user a favorable feeling for the robot and this feeling is enhanced when eye contact is used in combination with joint attention.

3 CONTRIBUTIONS

This work differs from the related work described above in several ways. First, concerning the activity of the interaction, instead of exclusively focusing the task in joint attention situations, the activity was designed to be a more natural and physical task that participants could accomplish even without the robot’s guidance. Second, regarding the robot’s physical appearance, we used a human-like robotic head capable of inducing and responding to joint attention directives through eye gaze, head pose direction, facial expressions and verbal cues. Furthermore, in order to track the participant’s behavior and the state of the task, our system relies on non-intrusive state-of-the-art technology to sense speech, gaze direction and moves played by the user.

Within the context of the tangible and natural interaction setup described above, we contribute to joint attention research in human-robot interaction by:

- Elaborating on the design of an easy to reproduce semi-autonomous wizard-of-Oz setup where a wizard, through an optimized interface, can take the perspective of the robot and perceive multimodal information in real-time;
- Evaluating user perceptions (from both first and third party perspectives) of an autonomous joint attention system that uses multimodal information to minimize the robot’s idle time and increase the usage of more meaningful joint attention behaviors.

4 MULTIMODAL JOINT ATTENTION SYSTEM

This section describes a semi-autonomous joint attention system developed with the intent of studying the effects of visual perspective taking and joint attention in human-robot interaction.

4.1 MagPuzzle task and Hardware

MagPuzzle is a spatial reasoning cooperative task where a life-like social robot interacts with human participants. Participants visualize a three-dimensional cube and reconstruct it by placing and manipulating six different colored puzzle pieces in a two-dimensional board with a 4X4 grid (see Figure 1). The task contains three different sub-tasks or puzzles, with varying levels of difficulty, where the maximum amount of pieces in a line (row or column) allowed for the solution varies. The social robot introduces the task and cooperates with participants to arrive at solutions in all three different puzzles. The robot guides users towards a solution mostly by using non-deterministic speech and deictic gaze behavior. The task was purposefully designed so that the robot’s guidance is more successful if joint attention is established with the robot, and so that the robot’s guidance is not essential for the completion of the task.
The hardware setting for MagPuzzle is comprised of two separate setups. A wizard-of-oz setup where an experimenter uses a dedicated computer and a screen equipped with a commercial gaze tracker from Tobii\(^1\) to control elements of the interaction, and the main setup located in an adjacent room where participants complete the MagPuzzle task alongside a robot. A back-projected robot head [3] is placed next to a table that contains the puzzle board. We decided to use a back-projected robotic head because they have been shown to be capable platforms for allowing gaze reading by humans as they allow for very fine grained deictic gestures [8]. Our robot is both capable of digitally animating quick eye gaze changes and also to use two degrees of freedom servos to simulate head movements. This is essential for a task such as ours, where we want users to assume the visual perspective of the robot. The robot height was adjusted to make the interaction comfortable and accommodating to most heights. In front of the robot, we are using an RGB-D camera that tracks the user’s gaze direction and position. For object detection, we are using a dedicated RGB camera on top of the board that detects fiducial markers placed on the pieces. For speech recognition and audio processing, we are using a high-quality amplified microphone placed on the side of the robot. One of the key aspects we focused while designing the hardware setup was to keep it simple and non-intrusive. In contrast to many studies in this area, users were not requested to wear glasses, participate in any calibration step or wear dedicated microphones. These decisions limit the performance of our multimodal perception system but do not influence the flow and naturalness of the experience.

4.2 Dialog Acts and Wizard Interface

Our system’s dialog acts are used to describe and manage the rules of the task, provide simple responses to users’ questions, offer advice on what to play next, provide feedback on a move that was just played and motivate or compliment the user. Each dialog act contains one or several behavior implementations and when a dialog act is selected one of its behavior implementations is chosen. The chosen behavior is parsed by a behavior planner that outputs the right multimodal output commands at the right time to the robot. During their duration, behaviors are displayed as subtitles in the wizard interface because the wizard does not have access to video or audio feeds of the interaction. We decided to restrict our wizard interface with displaying only the information that the robot can observe so that the wizard can more accurately take the perspective of the robot and decide when and how to act only with the information that the robot can also collect. The advantages of just exposing the wizard to what the robot can understand about the world have been previously shown [27]. Our behaviors are inspired by BML (Kopp, et al., 2006) and allow the mix of verbal and non-verbal behaviors in a format that is compact to author. The set of behaviors used in the task were authored to reflect a friendly cooperative social robot that often smiles, displays facial expressions calls participants by their names and can refer to puzzle pieces by their color.

The wizard can trigger dialog acts by clicking on buttons with the intended dialog act label. Labeled buttons can be placed directly on the interface or associated with expandable interface elements. The expandable interface elements are the 2D representation of the puzzle pieces, symbols of board quadrants, hint squares, state (in)correctness symbol, robot icon, and the player icon. The wizard can look at these interface elements to expose the buttons associated with them. Most of the times, the wizard intuitively looks at changes in interface elements (e.g., a new piece appears in the interface) and the associated buttons appear automatically. This confined natural exposure reduces the wizard’s cognitive load as only the options from the wizard’s current focus of attention are displayed on-screen. Gaze tracking is also used when the wizard quickly shifts his gaze to remote parts of the screen. In these cases, the mouse pointer will appear in the region of the screen the wizard is looking at to reduce response time. We also developed a helper algorithm that makes some interface elements appear only when relevant and disappears when no longer necessary. This algorithm tries to ensure that the wizard chooses the correct dialog acts at the right

---

\(^1\)https://www.tobii.com/
moment by continually monitoring the board state and all possible moves. By consulting the interface, the wizard automatically knows the correct place for providing hints to the user, if the board is in a proper or incorrect board state or if any rule was broken. The last important design decision for our system was to place all of the important multimodal elements of the interface close together so that the wizard is aware of all the changes in the interaction. Users can naturally communicate with our robot through speech, by changing their head orientation and by moving pieces on the board. All of these events are detected by our system and displayed to the wizard in real-time (see Figure 2).

4.3 Multimodal Perception

Speech recognition results are obtained from a commercial cloud service and shown to the wizard in a text box on the bottom-left of the wizard interface. Intermediate speech recognition results are presented with an ellipsis at the end of the detected utterance. Once the final result is known the punctuation is removed and the result stays in the interface until a new speech recognition event starts. In addition to this visual feedback, speech recognition results are also vocalized at two times speed in text-to-speech to the wizard’s computer. This audio feedback ensures that the wizard perceives what the user is saying even if focused on the other modalities of the interaction. We are using GazeSense, a non-intrusive gaze head direction tracker used to estimate participants’ visual focus of attention. This estimation is used to continuously display two semi-opaque lines that represent the user’s gaze. When the system detects that the user is gazing at a point of the board, or at the robot, the lines change, and the wizard becomes aware of where the user is looking. Given limitations in the performance of non-intrusive gaze systems we cannot distinguish individual squares in our board but rather four different quadrants each representing four squares (top-left, top-right, bottom-left, bottom-right). Using a single camera to precisely track the eye’s pupil proved not adequate for detecting gaze directed towards a horizontal surface. After performing pilot tests using our wizard-of-Oz system, we decided to only use the head orientation data as an input for our gaze estimation. Research has shown that head orientation contribution in estimating overall gaze direction is estimated to be around 70% [29] and relying solely on head pose to determine the VFOA of humans while interacting with a robot has been reported to lead to reasonable levels of accuracy in real-world real-time measures [18]. For tracking the puzzle pieces, we have integrated an open-source augmented reality toolkit that can robustly recognize fiducial markers placed on the puzzle pieces in real-time. When the system loses or gains tracking of a puzzle piece, the location is immediately updated in the center-left position of the wizard interface so that the wizard is always aware of the state of the board.

4.4 Autonomous Joint Attention System

The robot’s attention system produces gaze shifts (see Table 1) that are based both on real-time perception data and dialog act selection. When a dialog act is selected, the robot first shifts its focus of attention towards the intended target and only then executes the behavior. This behavior follows findings that referential gaze in speech precedes the onset of linguistic references [11]. If selected dialog acts are associated with a square on the board or a quadrant, the robot shifts its current focus of attention to the associated place on the board. In dialog acts not associated with the board, the robot changes its current attention target to the participant. The robot can precisely look at different parts of the board because it stands in a fixed position and gaze shifts towards board positions were manually calibrated. While focusing its attention towards a participant, the robot uses information from GazeSense to track the user in real-time at 30 frames per second. If the system loses track of the participant, it is assumed that the user is located right in front of the robot in a central location. When the behavior associated with the dialog act finishes, the robot performs an idle gaze shift.

Idle gaze shifts occur when a dialog acts finishes or when the timer for a previous gaze shift ends. In these cases, the system automatically triggers a new random gaze shift with a new maximum time duration. A study found that people dwell on referenced objects for approximately 1.9 seconds on average with no difference in the amount of time when the referencer was a virtual agent or human [25]. Also, gaze fixations that are too short might be interpreted as shyness or avoidance, whereas long gaze fixations may appear menacing or uncomfortable [1]. We try to avoid long gaze fixations by making our system react to most events in the environment and by randomizing gaze behavior after periods of inactivity. Based on these findings we decided to vary the robot’s idle gaze shifts between 1 and 3 seconds. The possible targets for an idle gaze behavior include looking at a player, at one of the four quadrants in the board, or at a direction that makes the robot seem that is either distracted or avoiding gaze. The probability of choosing each of these events is not equidistributed and was fine-tuned after extensive experimentation to produce natural looking behavior.

When not in the middle of a dialog act, the multimodal perception cues described in the previous subsection are also used to shape the autonomous joint attention system. The robot reacts to several events triggered by the participant and actively tries to respond to the participant’s focus of attention to establish joint attention. When a participant moves, places or removes a puzzle piece from the board, the robot assumes that the piece is the current focus of attention and gazes at that location. Human-human conversation partners often show their attention to each other by establishing mutual gaze and looking at each other when they are listening or speaking [4].

As such, when speech activity is detected from the user, the robot looks at the user to simulate that it is paying attention. Finally, the robot looks at the approximate positions of the board that the user is looking at or looks back at the player if the player is looking at the robot.

Gaze shifts triggered by perception events have priority limitations. The robot does not react to perception events if it is involved in a gaze shift with a higher priority. Unless a gaze shift with the same or superior priority is issued, the robot will maintain its gaze target until the time duration of the current gaze is completed and an idle gaze shift is issued. Within any shift of attention, the robot instantly moves its eyes and then follows the movement with the movement of the neck. This gaze behavior produces believable results and is default in our robotic platform.
5 EVALUATION

This section describes the methodology of the MagPuzzle study for testing the effects of visual perspective taking and overall success of the joint attention system, including an analysis of the objective and questionnaire results, followed by a third party observer study and respective findings.

5.1 Participants and Procedure

Twenty nine participants took part in a 30-45 minute between-subjects experiment in our [Anonymous] institute. Participants were mainly bachelor, master and Ph.D. students with ages ranging from 18 to 34 years. They were recruited via flyers placed on the campus and were rewarded with a cinema ticket for participating in the experiment. Technical issues and a change in protocol after the initial sessions reduced our participant pool to 22 (11 females, 11 males).

At the beginning of the experiment, participants were guided to a soundproof room separated from the wizard and were asked to fill a consent form and an initial questionnaire. Next, participants were briefly debriefed by the experimenter and instructed to sit down in front of the robot. At the beginning of each interaction, the robot explains the MagPuzzle task in detail and participants complete the set of three different puzzles in a time span that lasted from 5-10 minutes depending on the participant. Finally, participants fill in a final questionnaire that completes the experiment.

5.2 Manipulation

A separate condition that restricts the effects of responding to Joint Attention was developed. In this condition, the autonomous gaze system ignores the joint attention events from the participant actor to generate gaze shifts (see Table 1). Additionally, the wizard interface does not show the participant’s gaze direction. The wizard still initiates joint attention by choosing dialog acts that are either targeted at the board or the player, but in idle moments the randomized gaze will take over. As previously stated the timings for the idle gaze shifts were based on several trials and create believable patterns that go back and forth between looking at a position in the board, the player or looking distracted. This condition has the potential of being perceived as less mechanical and with an increased sense of agency but the robot’s behavior will likely look more detached from the user. To be able to discuss these hypotheses, in addition of testing the overall success of our joint attention system in the MagPuzzle task, we will separate our participants in two groups:

- **Control**, where the robot is only able to initiate joint attention;
- **Full joint attention**, where in addition to initiating joint attention the robot also accompanies and responds to the user’s attention.

5.3 Measures

In the initial questionnaire filled at the beginning of the experiment, participants report their demographic data, how often they are exposed to different technologies (robots, digital assistants and video games) and their personality through a reduced version of the big-five questionnaire[10]. The final questionnaire contained several five point Likert scale items that measured task difficulty, enjoyment, helpfulness, alliance and social presence. Difficulty was assessed by averaging three items that asked how hard the task was for the participant in each of the three rounds. For measuring enjoyment, helpfulness and alliance we picked a set of 8 different questions that were tested for reliability. For measuring social presence we are using a validated 36 item questionnaire [12] that measures the degree to which an individual feels interconnected with another entity [5]. To achieve a feeling of interconnection, participants should believe that both the robot’s behavioral and psychological involvement is connected to theirs and that their behavioral and psychological involvement is connected to the robot. The bidirectional nature of this measure makes it a sound choice to evaluate the effectiveness of our manipulation.

In addition to collecting questionnaire data, we video recorded the interactions and logged all of the information collected by the joint attention system. This information includes all gaze shifts, speech recognition results and moves from the users but also the gaze shifts produced by the robot, the dialog acts triggered by the wizard and its gaze information. All of this information will be used to objectively analyze the effectiveness of the proposed joint attention system. To some authors, joint attention is only established when participants are aware that they are establishing it. Since we cannot read the participant’s mind, similarly to [23], we consider the establishment of dyadic joint attention when participants are either involved in mutual gaze (looking at each other) or shared gaze (looking at the board at the same time). Using this definition, across conditions, we will compare the percentage of time that participants engaged in joint attention. This measure is also similar to the With-me-ness measure, a concept that aims to measure the extent that a human is with the robot over the course of an interactive task [18]. Finally, most of the hints provided by the robot during the task depend on the participant’s ability to adopt
the visual focus of attention of the robot and perceive the game from the robot’s perspective. We will analyze the effectiveness of our life-like social robot and joint attention system to provide hints by looking at the success rate of these cues.

5.4 Objective Results
After the analysis of the logged data obtained in the interactions, we observed that our robot could successfully evoke feelings of visual perspective taking because participants could correctly assess and act according to the cues exhibited by the robot at a high rate. The hints of where to play the next piece were given solely by the gaze and head pose direction of the robot (e.g., Putting a piece here might help) and, although this type of cue was difficult to perceive given the resolution of the board with pieces placed close together, the success rate of these hints was 65% which proves that participants had a good understanding of the robot’s viewpoint of the board. Nonetheless, the success rate of combined verbal and non-verbal cues correcting bad piece moves through color features and gaze direction (e.g., I would arrange this yellow piece differently) reached a 93% success rate.

For assessing the difference in the robot’s gaze behavior between the control and full joint attention conditions figure 3 is presented, illustrating the average robot’s gaze distribution with N=11 participants in each condition. A Student’s t-test indicated that the average robot’s gaze towards pieces was significantly greater in the full joint attention condition (M=.16, SD=.09) than in the control condition (M=.31, SD=.08); t(20)=−4.02, p=.001. On the other hand, for the average robot’s gaze towards the participant a significantly greater result was found in the control condition (M=.71, SD=.09) comparing to the full joint attention condition (M=.59, SD=.09); t(20)=3.08, p=.006. As expected, these results highlight the difference in the robot’s gaze behavior between both conditions and reflect the increased robot’s attention to the board (pieces and quadrants) in the full joint attention condition, which was intended in order to respond to the participant’s attention.

Finally, to investigate if our manipulation in fact increased joint attention between the participant and our robot, we employed a Student’s t-test on the results obtained for the average mismatched and shared attention, presented on figure 4 in percentage form. We considered N=5 participants from control condition and N=8 participants from the full joint attention condition. Indeed, even with the low number of participants included in the analysis (a consequence of problems in logging the data in the first sessions), we obtained borderline significant results for joint attention (both precise and imprecise). For precise joint attention, a borderline significant result was obtained in the full joint attention condition (M=9.57, SD=8.11) compared to the control condition (M=2.05, SD=2.12); t(11)=−2.00, p=.071 and the same pattern was observed for imprecise joint attention where we also obtained a borderline significant result in the full joint attention condition (M=12.94, SD=8.12) comparing to the control condition (M=5.53, SD=4.66). Hence, these results suggest that our manipulations were successful in the sense that our robot demonstrated more joint attention in the full joint attention condition. A lower percentage of attention mismatch would probably not feel as natural as the robot would look to mechanical and exaggerated in responsive behavior.

Figure 3: Pie chart showing the average rate of time of the robot’s gaze distribution.

Figure 4: Average percentage of time of mismatched and shared attention. Precise joint attention refers to when the user and the robot look at the same quadrant. Imprecise joint attention considers the entire board.

Figure 5: Mean values for our questionnaire measures. (SP User-Robot) total perceived social presence connection with the Robot; (SP Robot-User) total perceived robot’s connection with Participant.
5.5 Questionnaire Results

Cronbach’s $\alpha$ tests were performed to assess the scale reliability of the measures used in the study. The enjoyment scale consisted of 2 items ($\alpha = .941$), helpfulness consisted of 3 items ($\alpha = .915$) and the alliance scale consisted of 3 items ($\alpha = .816$). The social presence questionnaire was previously validated in [12]. For the first three scales we used Student’s t-test. No significant differences between conditions were found for the measures of enjoyment and helpfulness. Enjoyment appears to be inversely correlated with exposure to technology ($r(22) = -.455, p = .034$) and Helpfulness negatively correlated with how difficult participants considered the task ($r(22) = -.489, p = .021$). For alliance, we found borderline significant results when comparing the control (M = 4.18, SD=0.74) with the full joint attention condition (M = 4.67, SD = 0.39); $t(20) = -1.93, p = 0.068$.

For assessing differences between our conditions in the social presence scale and sub-scales, we employed non-parametric Mann Whitney-U tests as suggested by the original paper [12]. The scores for our measures are presented in Figure 5. A Mann-Whitney test indicated that the total perceived robot’s connection with participant was significantly higher in the full joint attention (Mdn = 3.6) than in the control condition (Mdn = 3.1) $U = 27.5, p = .030, r = .46$. There were no significant differences in the other direction of social presence when comparing both of our conditions. To better understand the cause of these results, we further analyzed the social presence sub-scales in both directions (see Figure 6). Message understanding ($U = 29.0, p = .037, r = .45$) and emotional understanding ($U = 33.0, p = .068, r = .39$) were greater in the full attention condition and were the most relevant sub-scales responsible for this effect. The most revealing item from our questionnaire that clearly shows and describes this effect is one of the items from message understanding, “My thoughts were clear to the robot” ($U = 12.0, p = .001, r = .70$).

5.6 Observer Results

The final step of our study was to evaluate whether third-party observers could perceive a difference between our full joint attention and control conditions. As the questionnaire analysis only showed significant difference in the robot’s perceived connection with the participant, we decided to strengthen our results and focus our third-party observer analysis on this direction. For this, we first created two videos, both showing the same actor interacting with the robot. In one video the robot was running the full joint attention model and in the other video the control model. The actor was responding in the same way in both videos. Both videos were approximately 1:50 minutes in length and displayed the exact same verbal behavior from both the human and the robot side. We then created a crowd-sourcing task and asked 40 participants to indicate whether they could perceive a difference in robot behavior between the two videos. Specifically, the participants were presented with seven questions of which the first four mirrored items of the social presence questionnaire. Co-presence and Message Understanding were not included because they represent constructs that are harder to evaluate from an external perspective. We decided to represent each dimension of the questionnaire by a single direct item, to adapt
to the context and limitations of a crowd sourcing task. We added three additional questions concerning whether participants could see a difference in how well the robot was collaborating with the human, following the progression of the task and how helpful the robot was towards its human interlocutor. We also gave participants the possibility to indicate that they could not perceive any difference. Results (summarized in Figure 7) confirm the same positive trend in the direction of the full joint attention condition when focusing on the robot’s behavior. Chi-square tests were performed and the items EU $\chi^2(2, N=40) = 12.95$, p < 0.01, BI $\chi^2(2, N=40) = 12.05$, p < 0.01 and help $\chi^2(2, N=40) = 7.85$, pp < 0.05 had significant differences.

![Third-Party Observer Ratings](image)

Figure 7: Frequency of third party observers’ opinions about whether our two conditions were similar or if either was more successful in one of the following items: (AA) attentional allocation; (EU) perceived emotional interdependence; (EU) perceived emotional understanding; (BI) perceived behavioral interdependence; (collab) better collaboration ability; (task_prog) more aware of task progress; (help) more helpful.

6 DISCUSSION AND CONCLUSIONS

We presented design decisions and considerations required to design an effective semi-autonomous system that simulates joint attention in human-robot interaction. This system allows a human wizard to both perceive the environment from a robot’s “shoes” but also enables participants to take the visual perspective of the robot. This system aims to reduce the wizard’s cognitive load, enables real-time visualization of multimodal features and allows the control of a social robot that engages users in a spatial reasoning problem solving task. When participants had difficulties within the task, the wizard was successful at providing useful hints at the right moment. These hints required the user to successfully decipher the robot’s attention as our humanoid robot only uses gaze direction for deictic referencing.

In addition to providing hints by initiating joint attention processes, our robot autonomously responded to the participants’ multimodal cues that reveal their attention. The perception of these cues enabled a reduction in the time that our robot was engaged in random idle behaviors and allowed the robot to exhibit more meaningful reactive gaze shifts. These non-verbal behaviors, when compared to a control condition, did not improve task performance, but caused participants to perceive the robot as more socially present. In our full joint attention condition, users reported that the robot was more capable of understanding their behavior and their psychological involvement in the task. This effect was extended to third party observers who were also able to observe similar differences between our manipulation. The increased sense of social presence also appears to contribute to a higher feeling of alliance in the task, a measure defined by an increased level of cooperation, trust in the robot and task awareness. These results show the importance of enabling robots with joint attention mechanisms that are not necessarily relevant for the completion of cooperative tasks. We believe these mechanisms will contribute for establishing and maintaining believable long-term interactions with robots that cooperate and sense the multimodality of the social world around them.

ACKNOWLEDGMENTS

The work is supported by the Anonymous Grant under Grant No.: XXXXXXXX and Anonymous Grant Program (http://www.anonymous.anon/anonymous).

REFERENCES


