UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

Shared Perception in Human-Robot Interaction

Investigating social perceptual mechanisms in humans and implementing Shared Perception skills in robots

by

Carlo Mazzola

Thesis submitted for the degree of *Doctor of Philosophy* (35° cycle)

January 2023

Dr. Alessandra Sciutti	Supervisor
Dr. Francesco Rea	Supervisor
Prof. Giulio Sandini	Supervisor
Prof. Paolo Massobrio	Head of the PhD program

Thesis Jury:

Prof. David Vernon, *Carnegie Mellon University Africa*	External examiner
Dr. Salvatore Anzalone, *Université Paris 8 Vincennes – Saint-Denis*	External examiner
Prof. Nicoletta Noceti, *University of Genoa*	Internal examiner

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

to Sara. . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Carlo Mazzola
March 2023

Acknowledgements

I would like to thank Alexander Mois Aroyo for his support in programming the robot iCub in the user experiment described in Chapter 3. I also acknowledge Maria Tsfasman, Anja Philippsen, Serge Thill, and Yukie Nagai for the fruitful collaboration resulting in the work described in Chapter 4. I would also like to extend my gratitude to Marta Romeo and Angelo Cangelosi for hosting me at the University of Manchester and for supporting me with the development of the model described in Chapter 5.

I owe so much to all my colleagues in RBCS and CONTACT! Thank you because everybody with its own background, abilities, and personality contributed to shape this wonderful experience in IIT. Finally, I wish to thank Alessandra Sciutti, Francesco Rea, and Giulio Sandini for supporting and encouraging me throughout this period, for their valuable and crucial suggestions, but above all for their strong belief in interdisciplinarity, which allowed me to embark on this journey.

Part of this thesis was supported by a Starting Grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. G.A. No 804388, wHiSPER.

Abstract

Interaction can be seen as a composition of perspectives: the integration of perceptions, intentions, and actions on the environment two or more agents share. For an interaction to be effective, each agent must be prone to "sharedness": being situated in a common environment, able to read what others express about their perspective, and ready to adjust one's own perspective accordingly. In this sense, effective interaction is supported by perceiving the environment jointly with others, a capability that in this research is called Shared Perception. Nonetheless, perception is a complex process that brings the observer receiving sensory inputs from the external world and interpreting them based on its own, previous experiences, predictions, and intentions. In addition, social interaction itself contributes to shaping what is perceived: others' attention, perspective, actions, and internal states may also be incorporated into perception. Thus, Shared perception reflects the observer's ability to integrate these three sources of information: the environment, the self, and other agents.

If Shared Perception is essential among humans, it is equally crucial for interaction with robots, which need social and cognitive abilities to interact with humans naturally and successfully. This research deals with Shared Perception within the context of Social Human-Robot Interaction (HRI) and involves an interdisciplinary approach. The two general axes of the thesis are the investigation of human perception while interacting with robots and the modeling of robot's perception while interacting with humans. Such two directions are outlined through three specific Research Objectives, whose achievements represent the contribution of this work. i) The formulation of a theoretical framework of Shared Perception in HRI valid for interpreting and developing different socio-perceptual mechanisms and abilities. ii) The investigation of Shared Perception in humans focusing on the perceptual mechanism of Context Dependency, and therefore exploring how social interaction affects the use of previous experience in human spatial perception. iii) The implementation of a deep-learning model for Addressee Estimation to foster robots' socio-perceptual skills through the awareness of others' behavior, as suggested in the Shared Perception framework.

To achieve the first Research Objective, several human socio-perceptual mechanisms are presented and interpreted in a unified account. This exposition parallels mechanisms elicited

by interaction with humans and humanoid robots and aims to build a framework valid to investigate human perception in the context of HRI. Based on the thought of D. Davidson and conceived as the integration of information coming from the environment, the self, and other agents, the idea of "triangulation" expresses the critical dynamics of Shared Perception. Also, it is proposed as the functional structure to support the implementation of socio-perceptual skills in robots. This general framework serves as a reference to fulfill the other two Research Objectives, which explore specific aspects of Shared Perception.

For what concerns the second Research Objective, the human perceptual mechanism of Context Dependency is investigated, for the first time, within social interaction. Human perception is based on unconscious inference, where sensory inputs integrate with prior information. This phenomenon helps in facing the uncertainty of the external world with predictions built upon previous experience. To investigate the effect of social interaction on such a mechanism, the iCub robot has been used as an experimental tool to create an interactive scenario with a controlled setting. A user study based on psychophysical methods, Bayesian modeling, and a neural network analysis of human results demonstrated that social interaction influenced Context Dependency so that when interacting with a social agent, humans rely less on their internal models and more on external stimuli. Such results are framed in Shared Perception and contribute to revealing the integration dynamics of the three sources of Shared Perception. The others' presence and social behavior (other agents) affect the balance between sensory inputs (environment) and personal history (self) in favor of the information shared with others, that is, the environment.

The third Research Objective consists of tackling the Addressee Estimation problem, i.e., understanding to whom a speaker is talking, to improve the iCub social behavior in multi-party interactions. Addressee Estimation can be considered a Shared Perception ability because it is achieved by using sensory information from the environment, internal representations of the agents' position, and, more importantly, the understanding of others' behavior. An architecture for Addressee Estimation is thus designed considering the integration process of Shared Perception (environment, self, other agents) and partially implemented with respect to the third element: the awareness of others' behavior. To achieve this, a hybrid deep-learning (CNN+LSTM) model is developed to estimate the speaker-robot relative placement of the addressee based on the non-verbal behavior of the speaker. Addressee Estimation abilities based on Shared Perception dynamics are aimed at improving multi-party HRI. Making robots aware of other agents' behavior towards the environment is the first crucial step for incorporating such information into the robot's perception and modeling Shared Perception.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Sociality is not exclusive to humans. In the animal kingdom, many species gather and live in groups. Yet, from an evolutionary perspective, sociality permeates human life so as to originate unique forms of cognition as well as collaborative, prosocial, and normative behaviors, up to the development of culture and morality [228]. The social environment shapes even perception. The awareness of others' attention, and perception, as well as the comprehension of their intention, thought and belief, is combined with sensory information received from the environment to perceive the external world. This way, the world is not perceived individually, and perception becomes *shared*.

Shared Perception occurs as the result of one observer's perceptual process in the presence of other social agents. A broader definition of 'social agent' allows studying Shared Perception also within the context of human-robot interaction (HRI). This is the perspective of the present research.

1.1 The interaction between humans and robots as the motivation of the research

Interaction can be thought of as the reciprocal action and influence established between two entities. Since perception is highly dependent on previous experience, personal perspective, and intentions toward the environment, humans perceive the external world differently from each other. Despite this, they can naturally and effectively interact. In everyday actions, passing an object or shaking hands, and even more in sports, dance, music, and complex collaborative activities, humans are very good at coordinating with each other. Regrettably, this is not the case in HRI. Robots still lack several abilities to autonomously interact with

the environment [239] and are even more impaired while interacting with humans. The quest for Shared Perception in HRI emerges here.

The situations in which embodied artificial systems are meant to interact with humans are numerous and different. Social autonomous robots interacting with humans as social agents represent only a part of them. Robots with social skills are developed, for instance, as companions in Therapy and Care Homes [39]. They have been proven to be beneficial as interactive tools to foster children's well-being [39] and education [21], also in the case of children with developmental disorders [194, 168]. Rehabilitation is another context where social robots have been demonstrated to provide positive effects in terms of perceived support and increased engagement in the rehabilitative task [20]. Socially assistive robotics may therefore benefit human society in different ways [59], but robots are also developed to populate other kinds of environments. Airports, banks, malls, restaurants, schools: these are some scenarios where social cognitive and perceptual skills become crucial to autonomously cope with human social dynamics.

Given this context, the present research is motivated by the need to improve the interactive social abilities of robots to make them instruments and collaborators of human development and well-being. The approach adopted for this problem is to start from perception because the ability to interact has a crucial point in the capability to perceive the environment jointly with others. Social interaction emerges from socially perceiving other agents and sharing the perception of the environment with them. Moving from the capability to perceive and interpret others' intentional relation towards the environment, Shared Perception would lead robots to understand their partner, augment the perception of the entire environment, and improve the quality of their social interactions.

1.2 The circle between investigating humans and implementing robots

If we aim for a natural and efficient HRI, the research should pursue two axes: deepening the knowledge of humans and fostering the development of robots. These are the two directions of this thesis and are tightly related, as in a circle, so that one supports the other and viceversa.

1.2.1 Pursuing research on humans to develop human-aware robots

Since humans are intentional agents rather than inanimate objects in the environment, to interact with them, robots need a model of human cognition and mind. Informed of this,

robots could understand human intentions, thoughts, and emotions implicitly expressed in their behavior. Interaction among humans is aided by the fact that we are of the same nature. Grown by interacting with other humans "like me" [144], we soon acquire social intelligence. For this reason, only by investigating the structure of human experience and achieving an extensive knowledge of cognitive mechanisms, it will be possible to develop human-aware robots and inspire novel approaches to improve robotic autonomy and social abilities [70].

There is also another reason motivating the interest in investigating humans for HRI. Interaction suggests reciprocity, and if robots should be enabled to understand humans, then the same thing would be desirable for humans: they should be able to understand robots. Therefore, moving forward knowledge of humans appears crucial to implementing human-inspired artificial models and making robots not only autonomous and efficient but also legible by humans [196].

1.2.2 Pursuing the development of robots to deepen knowledge of humans

Investigating human cognition provides inspiration for developing autonomous artificial systems. However, it is also true that implementing cognitive, artificial systems may provide novel, different insights to deepen knowledge of human experience. From this perspective, HRI is revealed as an interdisciplinary field of study. Multiple disciplines offer results and models that shine a light on human cognition and perception, but the research needs to be pursued. For instance, for what concerns the study of cognition during the interaction, only in the last decades, psychology, cognitive, and neuro-sciences are moving from an individual, passive approach toward a real interactive context [91, 24]. To study social cognition, humans need to be tested while interacting with other social agents. However, interaction with other humans does not always ensure the controlled and repeatable measures that cognitive and neuroscience require. The use of videos or virtual onscreen agents is not an optimal solution. Videos lose the possibility of reciprocity, making humans passive subjects, whereas virtual reality alters the natural environment of the interaction where cognition is normally exercised. HRI can help in solving this problem by providing a valuable interactive, embodied, and controlled setting employing humanoid robots as experimental tools [198]. Another possibility in this direction is offered by computational modeling of theories and reproducing cognitive phenomena. Also, the development of robots can be pursued with the same aim: fostering the knowledge of embodied cognition through the implementation of humanoid platforms (see, for instance, [147]).

1.3 Research Objectives

Connections between investigating human cognition and implementing cognitive embodied artificial agents become therefore evident in the form of a self-perpetuating circle: one direction sustaining the other. Human cognition inspires the development of robots and this latter may lead to extending the knowledge of human experience. From a general perspective, this work keeps the same overall orientation, but it focuses on the topic of Shared Perception and targets three specific Research Objectives (RO).

- **RO1: Formulating a theoretical framework of Shared Perception.** This objective consists of theorizing a general account of Shared Perception starting from a phenomenological perspective and pre-existent literature about human perception in social interaction and by structuring the theoretical account to be functional for robots' development.
- **RO2: Investigating the mechanism of Context Dependency in human visual perception of space during social interaction.** With this aim, this research deepens the study of one aspect of Shared Perception in humans by analyzing the impact of interaction with a social agent on a specific perceptual mechanism (Context Dependency) by using a humanoid robot to provide a controlled, embodied set-up.
- **RO3: Developing a model for Addressee Estimation to foster robots' socio-perceptual skills based on the Shared Perception framework**. This objective is made concrete by designing and implementing a deep-learning model to make the robot able to estimate the position of an utterance's addressee with respect to the speaker starting from visual information related to the speaker's non-verbal behavior.

1.4 Humans and Robots: two different observers, one general framework

If we generally consider Shared Perception as the observer's perception during social interaction, previous literature in HRI already addressed the topic either from the perspective of a human observer or from a robot's one.

In the first case, for instance, user studies employing HRI settings proved that social robots affect human perception. Robots could trigger phenomena such as perspective-taking [258, 128, 254], joint attention (for a review see [35]), or attribution of mental states (for

a review, see [224]). Similar mechanisms have been proven to be effective in establishing common ground between the two partners and helping humans in solving a collaborative task (e.g., see [139]).

Symmetrically, research strived to develop social robots endowing them with human-aware perceptual skills that, in a broad sense, could be considered related to Shared Perception. For instance, given the significant amount of information conveyed by gaze, the ability to compute the visual focus of attention has been implemented in robots with different techniques (e.g., see [201, 165]). Moreover, Waldhart et al. [242] developed a model for real-world HRI settings by employing a perspective-taking approach to improve the robot's ability to provide route directions. Robots endowed with perspective-taking abilities have been proven to be capable of enhancing human performance in an HRI collaborating task [53]. Intention reading is another ability that, for instance, Vinanzi et al. [241] developed, starting from human body posture and eye gaze direction to disambiguate goals in a collaborative task.

Considering the various perceptual phenomena and abilities connected to perception during social interaction, the present research addresses the topic starting from a third point of view: a framework for Shared Perception valid for both kinds of observers. The approach underlying the entire research proposes to tackle Shared Perception in HRI by theorizing a general account of Shared Perception, inspired by human social cognition, but functionally valid for the development of social robots. From this perspective, Shared perception is defined in Chapter 2 by introducing the concept of triangulation, inspired by D. Davidson [46]. With this concept, Shared Perception is outlined as the ability to integrate three different sources of information: perceptions of the external world (environment), internal models of reality (self), and perceptions of others as revealing their intentional relation toward the environment (others). The integration of three elements, i.e., the triangulation, is the dynamics of Shared Perception. Thanks to it, the observer can achieve an augmented perception of the environment not only by balancing sensory information (environment) with previous experiences (self) but also through the information from other agents that, therefore, becomes co-subjects in the observer's perception.

1.5 The structure of the thesis

Pursuing the first Research Objective (RO1), Chapter 2 aims to outline a theoretical framework for Shared Perception. Findings from empirical research on humans and several theories about perception are integrated together in view of a unified account of Shared Perception,

whose function is to provide a general overview and theoretical structure to interpret different human perceptual phenomena related to Shared Perception. Moreover, the framework is outlined to suggest a structure for implementing software architectures for social robots.

If Chapter 2 approaches Shared Perception from a general and theoretical perspective, Chapters 3 and 4 direct the focus on a precise human perceptual phenomenon. Context dependency is a well-known perceptual mechanism that had been studied only in individual scenarios. In line with the second Research Objective (RO2), the aim of these two chapters is thus to assess how interacting with a social agent (specifically a social humanoid robot) affects visual perception of space according to this particular phenomenon. Chapter 3 and 4 report two studies analyzing the same effect with different methodologies. Chapter 3 describes the procedure and results of a user study that reproduces the state-of-the-art setting to investigate Context Dependency in human perception but introducing an interactive scenario with a humanoid robot. Psychophysical methods, Bayesian modeling, HRI design, gaze analysis, and questionnaires have been applied to investigate this phenomenon. Chapter 4 aims to extend this research from a different angle. The phenomenon of Context Dependency can be interpreted in line with the predictive coding theory. Therefore, in this study, the participants' perceptual data recorded in the above-mentioned user study have been used to train an artificial neural network designed on predictive coding theory, to compare and deepen previous findings with a computational approach.

Chapter 5 targets the third Research Objective and tackles the implementation of Shared Perception for a selected socio-perceptual skill for robots: Addressee Estimation. As a comprehensive structure of perception during interaction, which can be referable to many different situations, Shared Perception is not implemented directly. Rather, the design of the model of Addressee Estimation is inspired by the Shared Perception framework as it is defined in Chapter 2. Addressee estimation, which is the ability to understand to whom a person is talking, is a crucial skill for social robots. Shared Perception can support such implementation so that it may enhance the robot's abilities to understand others, perceive the environment, and engage in more natural interactions.

For what concerns connections among parts, the framework of Chapter 2 serves as a basis to deepen the empirical investigation of a specific perceptual phenomenon (see Chapters 3 and 4) and to inspire the implementation of an interactive perceptual skill for social robots (see Chapter 5).

1.6 Shared Perception: a human-centered approach to HRI

If Shared Perception occurs in the eyes of the observer, which integrates different sources of perceptual information, its impact informs the entire interaction. From the individual perspective of the observer, Shared Perception could be thought of as an augmented perception. Thanks to perceiving, understanding, and integrating others' intentional relation toward the environment, the observer acquires additional viewpoints on it. As a result, a double, parallel, benefit is gained by the observer: enhanced awareness of others and enhanced awareness of the environment. Moreover, an additional advantage may be achievable from the perspective of the whole interaction. If both agents engage in Shared Perception, that would strengthen the relationship between the two parties and boost the interaction in being more natural and effective.

In fact, in HRI Shared Perception is a fundamental condition for the interaction to be enhanced, but still not sufficient. Beyond social perceptual skills, robots also need abilities to socially express their internal state. Without this, it would be difficult for humans to engage in Shared Perception with robots. In this research, this point is partially taken into account in Chapters 3 and 4, where the social behaviors of the iCub robot are varied in order to elicit Shared Perception in humans. However, this is a vast field of research that is mainly connected with robots' expressive abilities and with the idea of explainability in HRI through social cues [243].

Inspired by human cognition and aimed at improving interaction with humans, the present research examines Shared Perception with a human-centered approach from beginning to end. To make robots collaborators more than tools [192], the interaction with them needs to be designed on a human scale. Only this way, the user-friendliness, effectiveness, and fluidity of the interaction may be enhanced [196].

Chapter 2

A theoretical framework for Shared Perception in HRI

2.1 Introduction

To give a first formulation, Shared Perception can be thought of as the social ability of one observer to perceive the world by integrating information coming from the environment, the self, and other agents: more specifically, perceptions of the external world (environment), internal models of reality (self), and perceptions of others as revealing their intentional relation toward the environment (other agents). Starting from Human-Robot interaction, this chapter aims to outline a theoretical framework to understand what Shared Perception is in humans and offers a formal structure for its implementation in human-robot interaction.

In humans, Shared Perception is a socio-cognitive capability based on understanding others as different intentional agents. It has its ontogenetic roots in the social abilities acquired by children in their first months of life [226, 149]. Outlining a framework of Shared Perception for human-robot HRI entails that there are two agents of different natures from which Shared Perception should be considered: the human and the robot.

From this perspective, the steps followed to outline the theoretical framework will be the definition of the research context (see Section 2.2); the description of the perceptual process (see Section 2.3); the outline of the impact sociality has on human perception (see Section 2.4); the exposition of the basal pillars and the dynamic structure of shared perception (see Section 2.5.

2.2 The research context

The present research emerges at the intersection of three different areas of robotics: Cognitive Robotics, Social Robotics, and Human-Robot interaction. While these domains have distinct objectives and definitions, they share several aspects. Before considering their connections, it is worth examining them alone.

2.2.1 Human-Robot Interaction

As a general concept,**"interaction" can designate the reciprocal action established between two entities, which affect each other**. For instance, it may express the relationship between two living organisms or between an organism and the environment. It can also be linked to how humans relate to several technological artifacts, such as robots. In this case, reciprocity does not always occur. So, for such a relation to be considered "interaction", it is crucial the human at least believes the relation is reciprocal.

The interaction may be designed in different ways, depending on the channel of the interaction, which could be physical, spatial, verbal, or non-verbal.

Physical human-robot interaction refers to force-sensitive systems that adaptively and physically interact with humans in the physical world [89]. Robots for physical rehabilitation, wearable assistive robots, co-bots, or touch-based responsive robots all fall within this group. In these cases, the physical interaction is made possible by an exchange of forces or physiological electric impulses.

In spatial HRI [16], the channel of the interaction is related to the position the two agents hold in the environment. In this case, the reciprocal influence may imply avoiding others as obstacles or for safety reasons, or, on the contrary, a relation of guide-follower in the environment. Interactions based on human social proxemics are also included in this group.

Rather than being distinguished by the overt or implicit nature of communication, verbal and non-verbal interaction is classified by the code through which an addresser communicates a message to their addressee. Addresser, addressee, message, and code are four core elements composing the structure of communication according to Jakobson [105]. Given this preliminary context, and although verbal and non-verbal channels are strictly interlinked, they can be thought of as if verbal communication expresses its message through the syntax of a language. Conversely, non-verbal interactions convey messages through facial expressions, body movements, and proxemics [16]. Even para-verbal communication falls within the latter group for the meaning expressed by speed, pitch, inflection, and tone of voice.

2.2.2 Social Robotics

The focus on interaction recalls the theme of sociality. Nevertheless, it is worth noticing that not all interactions between humans and robots are designed at a social level. To clarify what Social Robotics investigates, it is necessary to consider what actually makes an interaction social. For instance, the interaction through the spatial channel is not social if the robot treats the human as a mere obstacle but may result social if the robot's behavior is driven, for instance, by proxemic reasons [156]. Hence, **the interaction can be termed social according to the extent the machine is designed to interact with humans based on their social nature**. But to do so, it is essential to understand how humans perceive robots in terms of their appearance and behavior. Taking as an example a physical human-robot interaction aimed at wrist rehabilitation, the mere exchange of forces between the machine and the human does not have any social connotation. Not even if the robot additionally displays performance measures to assist the human in the task. It may become social if the robot's appearance or behavior while communicating such indicators were designed following the cognitive/affective social nature of humans: encouragement, empathy, threat, reward, etc., therefore, producing a sense of engagement in the human [20].

To concentrate on the human perspective on interaction, a large part of research in social robotics focuses on studying how humans perceive robots, in terms of pure appearance and behavior. For example, several studies have demonstrated a correlation between the level of anthropomorphism of a robot and its acceptance by humans, which revealed the phenomenon of Uncanny Valley: the acceptance of the robot grows along with its human-likeness, up to a certain point, where the acceptance dramatically decrease if the robot is too anthropomorphic [152]. Other studies focused on the color of the robots [211], on their perceived gender [37], racial bias [2], or on how much they seem to convey agency, competence, intelligence, etc. [164]. Questionnaires often represent a suitable method to evaluate the human perception of robots. Still, the application of cognitive and neuroscience allowed going deeper into such investigation [248]. For instance, fMRI has shown that human neural correlates of vitality form recognition (gentle or rude arm movements) are the same in the case of perception of humans' or robots' actions [51]. Neural responses evidenced by hemodynamic signals through fNIRS have been used by [115] to determine the impact of human-robot eye contact on human social processing and quantify human social engagement with robots.

Another branch of social robotics is concerned with the robots' shape and behavior design. Taking inspiration from research on humans, the objective is to make robots appear more human-like, or closer to human expectations, preferences, and needs, but without entering the uncanny valley. Generally, the design of robots is strongly task-dependent. Androids [102],

such as Sophia and Geminoid HI-2, or half-bust robots, like Furhat [4], are more conceived as conversational robots, whereas Paro or Qoobo are designed with animal-like qualities for therapeutic companionship for older adults [203]. Nao, Kaspar, or Kiwi, are designed for child-robot interaction, sometimes specifically for assistance and companionship for autistic children [168, 21].

The aim of social robotics does not end with the design of the appearance and behavior of robots and how they are perceived by humans. It proceeds towards the implementation of socially interacting devices that can benefit human society in different ways [59]. Since one main worldwide issue is the high proportion of older adults, with a longer life expectancy and a strong impact on the quality of life caused by neurodegenerative diseases, social robots have been considered a possible solution. For instance, they have been proposed as companions for older adults with cognitive impairments, as service robots in elderly care (see [39] for a review), or as a tool for dementia screening to assess elderly people's cognitive functioning via social interaction [237]. Focusing on another age range, robots have been employed as a tool to foster children's well-being with different aims: emotional support after receiving a diagnosis, for dealing with disease, during the stay in hospitals, and distraction during medical procedures (see [39] for a review). Education is another field where robots demonstrated to be more effective than other more traditional learning technologies [21]. For instance, affective personalization was employed for 4-6 years old children to boost their engagement with the social robot and outcomes in early literacy education [166]. Verbal encouragement strategies used by a socially interactive robotic tutor enhanced children's learning of math [27]. A socially assistive child-robot interaction setting has been designed and proved effective in treating and improving severe dysgraphia [75]. Robots have been proven to be effective also in enhancing children's group dynamics by assisting inclusive processes [231, 79] . Eventually, robots have been demonstrated to be beneficial for children with autism spectrum disorder, to improve their social skills like joint attention in groups [194], enhance their performances in imitation tasks and use of language, and reduce repetitive and stereotyped behaviors [168].

2.2.3 Cognitive Robotics

As [193] pointed out the word cognitive has its etymological root in the concept of knowing, which among all living beings represents the most evolved way of interaction of an organism with the environment. Generally speaking, it is the passage from the sensory-based receptions of stimuli to their organization, realized through abstraction and allowing goal-based actions

toward the environment. Cognition relates, therefore, to the interaction of an organism with the environment aimed at adaptation and, as an ultimate end, at staying alive. Moving the idea of cognition to a generic autonomous system, either an organism or an artificial one, cognition could be defined as **"the process by which an autonomous system perceives its environment, learns from experience, anticipates the outcome of events, acts to pursue goals and adapts to changing circumstances"** [238].

Within this framework, the aim of cognitive robotics is the development of artificial models to provide robots with the ability to interact with the environment, learn, and adapt. To do this, a fundamental methodological issue concerns the degree of inspiration from natural systems. It can be argued that, since cognitive abilities primarily belong to the biological realm, every artificial cognitive model is somehow inspired by a biological system. Nevertheless, this does not mean that the implementation of such skills is biologically inspired as well. Most of them are based on engineering and heuristic principles, or are hybrid models, partly engineering/heuristic partly biology-inspired, [124].

In the case of biology-inspired models, other issues arise, related to what model is taken as inspiration. Not only human cognitive models are exploited: biomimetic robots inspired by animal behaviors represent an extensive field of research [72]. Also, among human-modeled robotics, it is worth considering which theory of cognition is used, what level of abstraction the model refers to (e.g., macroscopic vs microscopic representation), and which role is played by the body [238].

These issues define three features of cognitive robotics as a branch of research:

1. **The crucial role of interdisciplinarity**. Cognitive robotics is, therefore, at the crossroads of many different disciplines, so its aims are achievable only within a respectful and fruitful dialogue [31]. There is no single science approaching the study of cognition. Cognitive- and neuro-sciences, biology, chemistry, psychology, physiology, medicine, and philosophy all deal with it. Still, for the development and implementation of artificial models of cognition, other disciplines are also needed: at least, informatics, electronic-, mechanical-, and bio- engineering.

2. **The distinctiveness introduced by the role of the body**. Robots differ from other artificial systems primarily because of them being embodied. More precisely, for a cognitive embodied system, the embodiment does not only mean being implemented in physical form, a feature that belongs to most technological systems. It requires an active role of the body in the cognitive processes and the interaction with the environment [170]. In this way, for the artificial cognitive system, the body becomes the means for

grounded cognition, where processes linked to perception, action, learning, abstraction, self-awareness, etc. are all shaped based on the system's body.

3. **The inspiration circle between biological and artificial cognition**. Cognitive robotics is driven by a double motivation. The practical one is the building of artificial embodied cognitive systems; the principled one is gaining a better understanding of cognition [193]. Between the two levels, inspiration is a two-way process. The implemented robotic model is inspired by the biological one. Conversely, once implemented, the former may result in additional insights for the improvement of the biological model and a stronger theory of what cognition is.

Implementing an artificial cognitive system involves the development of modules for single processes and abilities (for instance, related to perception, motor control, decision-making, learning, meta-cognition, etc.) and, concurrently, the connection between all these modules. The software framework that integrates all these elements is called cognitive architecture [193]. More than one hundred cognitive architectures have been developed, each following different theoretical assumptions, methodology, structure, and technology [255]. As a possible taxonomy, cognitive architectures can be distinguished according to the type of information processing [54]. For instance, symbolic architectures base knowledge representation on symbols often organized along a set of if-then rules. By design, they are effective in planning and reasoning but lack flexibility [124]. Examples of this approach are ACT-R [181] or Soar [127] architectures. Instead of basing knowledge on symbolic entities, emergent models try to simulate processes of human cognition from the bottom up with artificial neural networks [255]. This allows the architecture to solve the problems of the symbolic approach, although losing transparency [124]. Examples are HTM and DAC. To address the issues of the two techniques, hybrid architectures seek to combine both approaches (as an example, see Clarion [218]).

Guided by the attempt to address the need for autonomy, adaptivity, and flexibility, and achieve greater biological realism, Vernon et al. [239] proposed a developmental approach for cognitive architectures in embodied robotic systems. Inspired by the emergentism of Maturana & Varela [140], this perspective is grounded on a concept of cognition that is not static, but under continual development, with a specific focus on cognitive development in infancy and childhood. Moreover, the authors propose this approach as a solution to facilitate meaningful social interaction between cognitive robots and humans, which is in the same direction as the present framework.

2.2.4 A shared field of interest

Although the perspectives of the three above-mentioned research areas are oriented in different directions, it appears evident that Human-Robot Interaction, Social Robotics, and Cognitive Robotics partially share their field of interest. The present research originates at their intersection. The interaction becomes the context of the research whereas the social and cognitive aspects modulate the relation between robots and humans. In this way, the interaction is built and studied according to the social dimension and cognitive frame in which the robot is developed. For humans, social interaction has been proposed as the default mode of the brain [90] and the base for the development of high forms of cognitive representations, enabling for instance metaphors, dialogic and reflective thinking [227]. Following this idea, social interaction should be considered equally fundamental when it comes to the implementation of artificial cognitive agents. Robots operating in human-populated, and thus social, environments need social skills. Furthermore, their cognitive processes (perception, action, decision-making, reasoning, etc.) should be shaped and strengthened based on social interactions. The intersection among the three branches can also be seen from another perspective. Social interaction has been defined as "two or more autonomous agents co-regulating their coupling with the effect that their autonomy is not destroyed, and their relational dynamics acquire an autonomy of their own" [47]. Following the same authors, social interaction is qualitatively defined by the engagement between the two agents [47]. In this respect, sociality and cognition improve human-robot interactions. The social dimension of the coupling offers a better mutual understanding between the agents, which may culminate in engagement, a parameter that is crucial to measure to evaluate human-robot interaction [5]. Moreover, advanced cognitive skills enable factual autonomy for robots, and the stronger the robot's autonomy, the greater the level of mutuality with the human. Three elements, therefore, delimit the context where the framework of Shared Perception rises:

1. Interaction: reciprocal influence between two or more social/cognitive agents sharing the same environment.

2. Sociality: relational modality based on the cognitive/affective social nature of humans. It is enabled by the way agents perceive each other and enables mutual understanding and engagement.

3. Cognition: the process that allows agents to perceive, make inferences, predict, act, adapt to the environment, and interact with each other.

2.3 The process of perception: from senses to the phenomenal object

The present paragraph does not aim to provide an exhaustive examination of all the accounts, paradigms, and theories of perception. Rather, it intends to briefly offer an overview of the perceptual process and, afterward, present perception from a phenomenological perspective. One of the points where most theories of perception agree is that there is a difference between the object as it exists in the physical world and the phenomenal object, as it is perceived by the observer. Perception is the process that, gives unity and significance to sensory stimuli and is thus responsible for the constitution of the phenomenal object. The difference between phenomenal and physical objects appears with visual illusions. For instance, there are cases where the phenomenal object is not perceived although the physical stimulus is given. On the contrary, in other cases, the phenomenal object appears without its physical correspondence being present (see Figure 2.1).

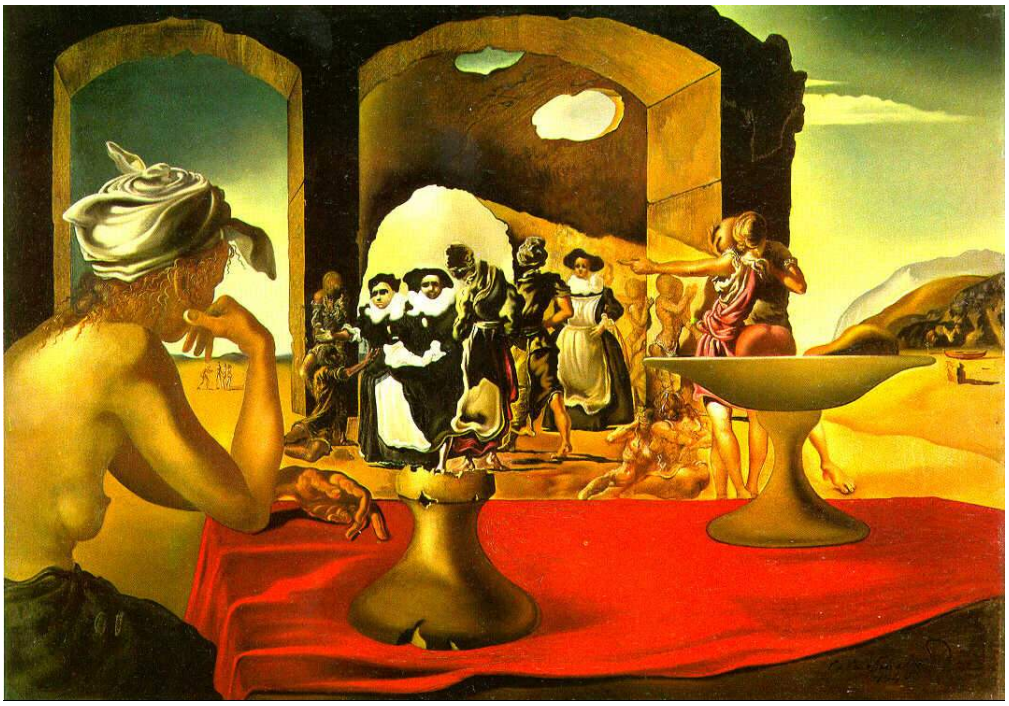

Figure 2.1 **Slave Market with the Disappearing Bust of Voltaire (1940), Salvador Dali.** This painting shows the blurred relation between the phenomenal object and the physical stimulus from which the bust of Voltaire appears. Collection of The Dalí Museum, St. Petersburg, FL (USA); Gift of A. Reynolds Eleanor Morse. ©Salvador Dalí. Fundació Gala-Salvador Dalí (Artists Rights Society), 2022.

As in a movement from sensory receptors toward the organization and unity of the phenomenal object, stimuli are processed in different directions. In the bottom-up direction, sensory stimuli are processed to recognize the fundamental components and properties of the object (e.g., regarding vision, colors, elementary shapes, and size). On the contrary, top-down processing is guided by previous experience, motivations, expectations, learning, etc. The context in which the physical object appears is an example of expectation exerting its influence on the phenomenal object. The two movements, bottom-up and top-down, are coordinated, integrated, and together lead to the constitution of the phenomenal object: the object as we perceive it.

2.3.1 The organization of sensations

Although the bottom-up and the top-down processes run together, the theory of Gestalt, which dates back to the first decades of the XX century, asserts the view of the overall shape (Gestalt) is prominent with respect to one of the single elements composing an object. According to Gestalt theory, the phenomenal object is not the result of the association of elementary sensations. Sensations appear as organized in a configuration, or pattern. Following this perspective, the top-down processing led by the shape guides the bottom-up composition of the object rather than the opposite.

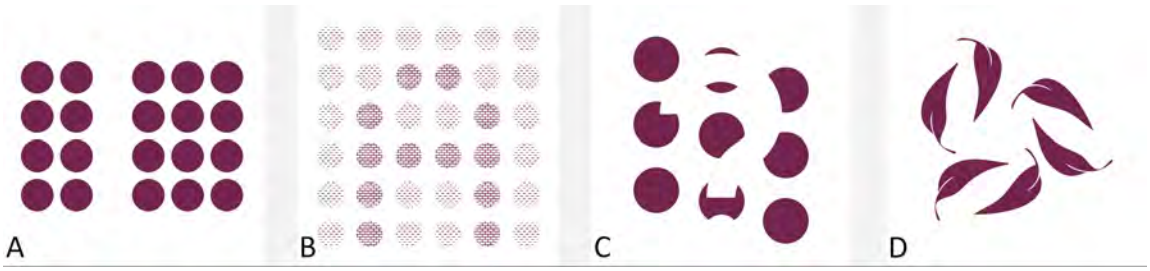

Figure 2.2 **Illustrations of four laws of Gestalt.** The principle of *proximity* affirms that elements close to one another in time or space tend to be perceived with a sense of togetherness rather than as singular elements. In Figure A, circles appear as unified in two groups. *Similarity* states that we tend to configure elements that share similar characteristics as a group, as it happens in Figure B, where we can distinguish the letter A only from the different pattern of some of the circles. The principle of *closure* indicates that lines tend to be unified in seeking a single, recognizable pattern, as is evident from Figure C, where a question mark appears from the circles. *Symmetry* affirms that we tend to perceive symmetrical elements together: in Figure D, it is intuitive to consider the group of 6 leaves as formed by three couples of symmetrical leaves.

Over the years, many principles of how information is organized have been identified. Some of them have already been proposed by the first exponents of the current, such as Köhler [121], Koffka [125], and Wertheimer [247]. For instance, the principle of proximity, similarity, closure, and symmetry, are represented and explained in Figure 2.2. One of the effects of these principles is related to the integration and segregation of information that underlies the difference between ground and figure. In Gestalt theory, Figure and Ground always go together since there is no one term without the other [188], an idea that is clearly visible, for instance, from the face-vase visual illusion in Figure 2.3.

Figure 2.3 **Rubin's vase illusion.** Figures and ground are inherently tight together. We either see the two white profiles of the iCub robot or the black vase: we cannot see the two things at the same time because the ground is necessary to see a figure.

2.3.2 The influence of experience

A different perspective is by H. Helmholtz, who theorized elementary sensations are synthesized in the perception of objects through association. From this perspective, he formulated the concept of "unconscious inference'" to recall the role of previous experience in generating perception from sensation [92]. In this case, the term "unconscious" does not refer to an innate or inaccessible level of the mental functions but to the automaticity of a process of which we are not consciously aware [232].

The idea of unconscious inference has been recently evoked to explain the integration of previous experience with sensory stimuli leading to the effect of "central tendency" [38, 113, 195]. This phenomenon, also known with the name of "regression to the mean" or "context dependency", dates back to Hollingworth, who in 1910 wrote: "Judgments of time, weight, force, brightness, extent of movement, length, area, size of angles, have all shown

the same tendency to gravitate toward a mean magnitude, the result being that stimuli above that point in the objective scale were underestimated and stimuli below overestimated" [93]. According to recent literature, the experience gained from exposure to previous stimuli is integrated with sensory information of the current stimulus and leads to the final perception: a process that, following a Bayesian account, can be described as an inference of previous experience [109, 38, 169]. This perspective is also in line with another theory of perception taking inspiration from Helmholtz: predictive coding. From this perspective, the inferential process is led by the prediction of the upcoming stimulus, which reflects previous experience, on the sensory information of the upcoming stimulus [202]. According to predictive coding, the brain functions as a prediction machine that continuously tends to match upcoming sensory information with expectations of the external world. These predictions operate in a top-down direction, and the resulting perception is ruled by a mechanism of minimization of the prediction error [40, 41].

2.3.3 The action-perception cycle

Rather than merely being passive observers, we continuously perceive the environment and act accordingly on it. Even more, our actions shape the way we perceive the world. The methods we employ to face our daily interactions with the world involve our bodies in a combination of perception and action [236]. Since childhood, our bodily experience has been the background of every learning process [58, 138]. For newborns and children, the development of action functionality impacts the acquisition of better perceptual skills. For instance, in the first two years of life auditory localization skills, and specifically the minimum audible angle necessary to localize a sound, improve more rapidly the more control infants acquire over their head movements [22]. From the same developmental perspective, the more precise children's motor control over hands and fingers, the more difficult objects' properties they can discriminate, e.g., size, texture, temperature, weight, shape, etc. [29].

"Active inference" is a theory that explains the strict relation between perception and action with the same approach of Predictive coding. A brain is a predictive machine aimed at reducing prediction errors. Hence, to achieve its aim, its first solution is to find the prediction fitting better reality, whereas the second is to perform actions that make predictions come true [41, 67, 171]. Following this account, "perceptual and motor systems should not be regarded as separate but instead as a single active inference machine that tries to predict its sensory inputs in all domains" [1].

2.3.4 The phenomenological approach to perception

The perceptual organization of sensations, the role of previous experience in modifying perception, and the reciprocal influence and relation between perception and action highlight the difference between physical and phenomenal objects. From another approach, not empirical but phenomenological, the issue of the perceptual object is addressed differently. By defining perception as an intentional act, Husserl underlines that

1. perceiving means always perceiving *something*
2. the perceived object is the same despite the different conditions and perspectives in which it may appear, that is, despite the different sensations of it.

For perception to be an intentional act means therefore that what is perceived is the object constituted (apperceived) from the sensations in which it appears and that perception keeps a direct reference toward the object. Following an example of Husserl, I can see a box, not my sensations [97].

Merleau-Ponty is another phenomenologist who devoted a significant part of his research to perception. He advanced the thought of Husserl bringing attention to the issue of the body. The intentional relation to what we perceive is realized through the body, which is not a mere instrument we can divest. As humans, we can act upon the world and perceive it, only as being a body. The body is our anchor to the world and being a body means not only that we are in space and time. Rather, we inhabit space and time. Motor experience is our way to get access to the world, that is, to the objects we perceive [146], and body schema is the non-conscious constant organization responsible for our body's operative performance in the environment [71].

A further step in the phenomenological research on perception was made by Richir. He focused specifically on the process that brings toward the perception of something. For Richir, the perceived object becomes present to the subject after a process in which its meaning is fixed and constituted, that is, takes a defined form. The process has its original roots in a flow, which Richir called "phantasia": a flow of appearances that occur before any semantic fixation of meaning. The first fixation of meaning happens with a second stream of images, memories, bodily affections, and predictions that, eventually, contribute to the constitution of the meaning of the object [179].

As a result of this analysis, considering the phenomenological approach and the findings from the different perceptual and cognitive accounts above-mentioned, it is possible to provide an operational description of the perceptual process. It can be defined as the process

that starts from sensations and brings to the constitution of a defined perceptual object under the influence of attention, memory, motor schema, emotions, predictions, and other representations. The concept of Shared Perception, as presented at the beginning of this chapter (the ability to integrate perceptions of the environment, internal models of reality, and perceptions of others), fits with the definition of the perceptual process. From a Shared Perception perspective, representations of the world, the self, and others are integrated together in the process, producing a unique perceptual object which condenses the meaning of its founding representations.

2.4 Perception and social cognition

When perceiving others, we tend to integrate into our perception their behavior toward the environment: that is their intentional relation to it. The first aim of this section is to describe some of those mechanisms and social abilities that are involved in the process of Shared Perception. Moll & Meltzoff [149] propose that Joint-attention, Perspective-taking, and Theory of Mind attribution are linked in child growth, as distinct stages of ontogenetic development. Albeit their ontogenetic connection, they represent distinct social abilities, involving different aspects of cognition. For this reason, they will be discussed separately in this Section. As a second aim, this Section wants to show that similar effects can be elicited in humans during interaction both with other humans and with robots.

2.4.1 Integration of the other's attention

Butterworth [30] defined "Gaze following" as "looking where someone else is looking", an ability that represents a primary step toward the development of higher social skills, such as "Joint attention". With respect to the former, the latter can be thought of as the co-orientation established in triadic interactions, which are interactions among the self, another agent, and a third element. More specifically, in Joint attention, an agent is oriented by the other's gaze toward a third entity (might it be an object, an event, or a person) with the mutual knowledge that the other is being attentive to the same entity (see Figure 2.4) [62, 229].

Mundy et al. [153] suggested infants' joint attention behaviors should be distinguished into two groups: responses to others' cues or spontaneous initiations. While the responding joint attention behavior is also present in Chimpanzees, different is the case for initiating behaviors, which seems unique for the human species [229]. At a neural level, the difference seems to be paralleled by two different neural systems [154]. The first would be in common

Figure 2.4 **Integration of the other's attention.** An example of human-robot joint attention elicited by the human. The robot's attention is attracted by the other's attention and is brought to detect the toy train on the table.

with primates, the second, which integrates internal monitoring of one's own motor control and goal-directed behaviors and external monitoring of others, would be exclusive of humans and develop starting between 4 and 6 months of age [154]. Studies highlighted the range between 9 and 15 months as a crucial age for the development of Joint attention behaviors [229]. At the 9th month, infants start understanding other individuals as intentional agents, i.e., pursuing goals, and thus engaging in triadic interactions. Afterward, starting from 10-11 months, they start showing correct inhibition of responding behavior when the caregiver looks at objects with closed eyes, revealing sufficient awareness of the referential intent of the caregivers' gaze [25]. Following Carpenter et al. [34], the development of joint attention abilities covers two steps. The first is to "understand that others have some kind of psychological stance that is different from our own". Whereas the second is to understand what their psychological stance consists of.

Social robots can establish indirect communication based on joint attention with their human partners. An experiment with a stuffed-toy robot revealed that the gazing behavior of the robot was effective in gaining the partners' interest and drew them to look at the same objects as the robot [256]. Following this perspective, the recent focus on social cognition and effects induced by intersubjectivity calls for ecological experimental set-ups [24] and social robots might meet this need [35]. Social robots ensure both experimental validity and ecological settings. They have been used as tools to elicit and study joint attention mechanisms (for a review, see [35]). The study and improvement of joint attention mechanisms in children with ASD [244, 6] represents one research area in which they have been employed. Although humans demonstrate to follow the gaze of robots, some differences still occur between engaging in joint attention with other humans or robots (for a review, see

[3]). Even so, this disparity has been demonstrated as strongly dependent on the sociality, behavior, and appearance of the robot [3]. When interacting with a robot, infants of 18 months who had already seen the robot behaving socially, were more prone to follow its gaze than those who has not had that experience [145]. Furthermore, it has been demonstrated that for a humanoid social robot to establish joint attention, it is essential to establish eye contact with the partner previously [122].

2.4.2 Integration of the other's perspective

Perspective-taking generally refers to the ability to grasp the thoughts, perceptions, and intentions of others. In this sense, it is theoretically grounded on Piaget's research about the development of children's ability to coordinate different perspectives [174]. In visual and spatial perspective-taking, the phenomenon has been defined as the ability to represent the others' viewpoint (i.e., what/how they see things) or where things are located with respect to others [219]. Furthermore, to investigate the nature of cognitive mechanisms leading to these representations, an important distinction has been made between Levels 1 and 2 of perspective taking [63, 64], where Level 1 refers to what others represent and Level 2 to how they represent it. Therefore, regarding visual perspective taking (VPT), Level 1 involves what another person sees, whereas Level 2 how they see it. As a further distinction between the two levels, Apperly & Butterfill [7] suggest that Level 1 results from an implicit mentalizing system that develops early and processes visual information automatically and fast. On the contrary, Level 2 would be led by a more complex, later developed, and flexible system.

A classical experimental scenario to measure Level 1 VPT involves the Dot perspective task. Participants are usually seated in front of a screen, and the task consists of counting the number of dots in the scene where also an avatar is present. VPT is demonstrated by the confounding factor of the avatar's perspective, which leads participants to respond slower in case it differs from their own. Different results are provoked by variations in the tasks, e.g., explicit/implicit indications to follow the avatar's perspective, occluded view for avatars, or directional objects (arrows) with no agency replacing avatars. As a consequence, a debate was raised concerning whether VPT was triggered by automatic, stimulus-driven mentalizing processes (e.g., [220, 68]), or sub-mentalizing, attention-orienting mechanisms (e.g., [42, 73]).

To propose another approach, O'Grady et al. [163] explain apparent irreconcilable previous findings as a result of a different experimental design. Providing further evidence to the literature, they reject directional orienting and automatic stimulus-driven mechanisms

as hypothetical processes underlying perspective-taking. Rather, they suggest VPT level 1 effects could be explained by spontaneous, involuntary cognitive processes dependent on attention and, more specifically, on how much the avatar's perspective is salient. The research of Zhao et al. [257] might be interpreted in line with this account. Authors investigated Level 2 VPT with the use of numbers (e.g., 6 and 9) that could be read differently according to the perspective of the subject or a second agent in the scene (see Figure for an illustration 2.5). They found that VPT effects are dependent on the strength of the trigger revealing others' perspectives: in ascending order the mere presence, the object-directed gaze, and the goal-directed action toward the object.

Figure 2.5 **Integration of the other's perspective.** An example of a perspective-taking scenario. The number is perceived differently by the two agents (either 6 or 9) but each agent is influenced by *how* the other perceives it (II level perspective-taking).

A similar experiment has been carried out to investigate whether a robot (Nao or Baxter) may elicit VPT effects like those elicited by humans [258]. As in the previous study [257], authors noted that goal-directed actions provoke stronger effects than object-directed gaze and simple presence in the scene. Furthermore, the effect is increased if the stimulus shown to participants consists of a video rather than a picture and is not different between the two robots employed in the study, although never reaching the level of human-human interaction.

Applied developmental research employed social robots to strengthen perspective-taking abilities in children with typical development [254, 253]. For instance, children with typical development were asked to guide a Cozmo robot considering its point of view, a task that consisted of improving their Level-2 Spatial perspective-taking skill [254]. The study demonstrated that the more children carried on in the task, the fewer were their errors in guiding the robot. In another study, children's task consisted of instructing a NAO robot to perform different game actions [253]. Although starting using an egocentric perspective, children learned to change perspective and used the robot's one, in particular, after the robot

failed the first actions. As for joint attention, social robots have also been used to enhance the social abilities of children with ASD, through the development of VPT. With the aim of aiding children to see the world from the robot's perspective. After creating an experimental protocol with different interactive games and pre and post-assessment tests [250], Kaspar robot has been used to aid children see the world from its perspective, with beneficial effects on their VPT abilities [128].

2.4.3 Integration of the other's actions

Besides others' attention or perspective on the environment, also their actions toward it may affect one's perception. As humans, we are able to interpret others' inner states from how they behave. Others' actions toward an object may reveal their intentions, beliefs, and affective states about it, as well as some of its properties (see Figure 2.6). For instance, Kaiser et al. [112] demonstrated that 5 to 7 years old children were already able to extract relative weight information from merely observing videos in which an actor was lifting and carrying a box. The perception of an actor grasping an object improves the accuracy of judgments about the object's size [85]. Evidence of others' actions influencing an observer's perception of the environment also comes from studies related to the ecological account of affordances [78]. According to the theory of affordances, humans do not perceive objects starting from their physical properties, e.g., color, texture, shape, and weight, but for their functionality and affordability to use, i.e., using a term coined by Gibson, for their affordances. Interestingly, this functional perception of the environment may also be directed to others' affordances (for a review see [44]), so humans perceive the environment by scaling its features to the action capabilities of other actors [215].

The ability to understand other's actions finds a neuro-physiological explanation in the mirror neuron system, a cerebral organization of neurons that activates both when one is performing an action or when they are merely seeing others performing the same action [183, 182]. In this way, the mirror neuron system contributes to action understanding by activating one's motor representation [61], implicitly mapping others' actions onto one's motor schema. Interestingly, it has been demonstrated that the mirror neuron system is not only involved in action understanding from the perspective of action recognition, i.e., what action is performed. It also activates providing information about actors' intentions, i.e., why the action is performed, grasping a cup either to drink or to clean [99]; and about the action's vitality form, i.e., how it is performed, passing a cup either rudely or gently [50, 49].

Figure 2.6 **Integration of the other's action.** An example of how the action of others can be integrated and used to perceive an object. The human is giving the stuffed green dragon *gently and carefully* to the robot, with two hands. A robot endowed with social perceptual abilities should perceive the stuffed dragon as something delicate or precious and, accordingly, use two hands to handle it.

The investigation of human cognition with humanoid robots led to interesting results about the influence of robots' actions on human perception-action processes. For instance, results from Kaiser et al. [112] about weight information extraction from human motion have been reproduced by Sciutti et al. [197]. In the latter paper, human participants' performances in inferring the weight of an object from either human actor's or robot's lifting actions were comparable. Another study investigating humans reading hidden agents' intentions and object properties from robot's action is by Lastrico et al. [131]. Here, a humanoid iCub and a Baxter robot were recorded while grasping, lifting, and handing to participants an object to communicate a careful or not-careful movement with velocity profiles replicating humans' kinematics [74]. The authors found that the different robot's motions were correctly perceived by participants and, furthermore, elicited motor adaptation in subsequent object manipulations.

From a neurobiological perspective, robots' actions have been demonstrated to trigger mirror neuron systems like humans [76]. Similar evidence was also found from EEG measurement, both in the case of actions object-directed or not-object-directed [160]. Not only, but human brain activations triggered by robotic movements replicating gentle and rude action styles did not differ from those triggered by human motion with the same vitality form [51].

2.4.4 Integration of the others' (attributed) inner states

Bottom-up processes are not the only way to come up with action understanding. On the contrary, it has been argued that even in the case of simple sensory stimuli, these models are not sufficient and require a top-down component [10, 40]. The other way for action understanding and behavior interpretation consists of mentalization, a cognitive ability that leads humans to attribute inner states to others. From a developmental point of view, this capability has been hypothesized to originate from more primitive social skills. Early, in the first months of life, a process bringing newborns from understanding that other persons are 'like me' to the development of social cognitive skills starts [144]. From the ninth month, infants start considering others as other intentional agents, an understanding they will gradually develop over the years [226]. Social development has its foundation also in the human-unique motivation of sharing psychological states with others. This typical tendency leads children gradually share behavior, emotions, goals, and perceptions and thus collaborate with others [229].

Gradual progress brings infants from joint attention capabilities to perspective-taking [149]. In early childhood, this process reaches a critical step. The children start distinguishing their mental states from those of others, avoiding the mere egocentric attribution of their mental states, hence ascribing mental states to others [149, 226], a capability defined as "Theory of Mind" [175]. More specifically, it refers to the ascription of others' inner states such as beliefs, intentions, desires, attitudes, and feelings about a target reference (see Figure 2.7).

Figure 2.7 **Integration of the others' (attributed) inner states**. An illustration of how others' inner states can influence the observer's perception of the environment. A social robot should infer something serious, dangerous, or undesirable just happened on the screen thanks to understanding the other's inner states and attention.

Attribution of inner states has been hypothesized to act as a top-down modulator of perception/action processes (for a review, see [223, 10]). For instance, it has been demonstrated that inferences about others' intentions (reaching or withdrawing) are incorporated into the perception of the action movement (disappearance of the movement) and that such inferences are due to explicit knowledge about the actor's intention and implicit information from action kinematics [96].

The automatic tendency to follow another's gaze is also modulated by Theory of Mind. In one experiment, participants watched videos of people believing to interact with them. People in videos wore a pair of glasses which were always identical, but in one condition participants were told that glasses allowed vision whereas in another condition they did not. In spite of an identical bottom-up stimulus, the attribution of "seeing" or "not-seeing" mental states affected the automatic response to gaze-following [222]. Attribution of mental states also affects the perception/action mechanisms. The same action was differently processed and hence automatically imitated when believed to be intentionally or unintentionally performed [134]. Therefore, these findings suggest that inner states attributed to others due to their behavior, action or contextual information have a role not only when collaborating with them but also in the perception of the environment shared with them.

The research on mental states attribution to robots by humans produced extensive results over the years (for a review, see [224]). In general, it has been demonstrated that humans ascribe internal states to robots. However, as Banks [12] shows, humans tend to implicitly mentalize robots if their behaviors are coherent with those expressed by humans. Moreover, the implicit ascription does not always correspond to the explicit one. Lee et al. [132] demonstrated that adapting the robot's nonverbal behavior to the participant's internal states led the participant to ascribe mental states to the robot.

Regarding the impact of inner states attribution on perception-action processes, it has been demonstrated that the ascription of mental states to humans and robots affected the bottom-up mechanism of gaze cueing. Independently of whether the trial face belonged to a human or a robot, the gaze cueing effect resulted larger if participants believed the gaze was controlled by a human [249]. Similar evidence resulted from using the event-related potential (ERP) technique. This revealed that even at a neurobiological level human attention was influenced by the human or the robotic gaze only if it was believed to be guided by a human mind, hence if internal states were ascribed to the other agent [251].

2.5 Building shared perception

The research context (Section 2.2), the description of the perceptual process (Section 2.3), and the outline of social mechanisms involved in perception (Section 2.4) provide the foundation for an account of Shared Perception, but to complete the framework and outline the structure of Shared Perception another step is needed. To this aim, Section 2.5.1 define the core pillars of the framework. Thereafter, in Section 2.5.2 the structure of Shared Perception is specified as triangulated dynamics among three elements: the self (of the observer), the other, and the environment.

Pillars give stability, define the skeleton of architectural structures, and allow the construction of buildings. The same function is performed by the core elements listed here: Embodiment (2.5.1.1), Intentionality (2.5.1.2), Human-Awareness (2.5.1.3), Hermeneutics (2.5.1.4), and Sharedness (2.5.1.5). These fundamentals let Shared Perception emerges between the (at least) two agents (observer and other) and the external world (environment). The pillars and the structure of Shared Perception are outlined inspired by human perception. For this reason, this framework could be considered a simplified, functional model of the human perceptual process during social interaction. Since it is contextualized within HRI, the role of "observer" and "other" are envisaged so that they can be applied either to robots or humans.

2.5.1 The Basal Pillars of the framework

2.5.1.1 Embodiment

A noticeable difference exists between perceiving the environment and perceiving other agents: the others are not mere objects in the environment. They are animated bodies situated in the environment. At least for humans, this means that the body cannot be worn and abandoned [146]. But this is also true for robots, in that the body identifies that robot (see Figure 2.8). In addition, the meaning of embodiment is not only associated with the agent's identity. Being a body also indicates that the body is the means, the center, and the basis of experience [146]. The body allows therefore to experience the world and is the direct expression of internal states. Internal states appear through the body and on the body, which does not mean they are entirely revealed on the bodily surface. Rather, their emergence and expression are embodied. The embodiment becomes the means that makes any interaction possible because the observer can understand others only through their body.

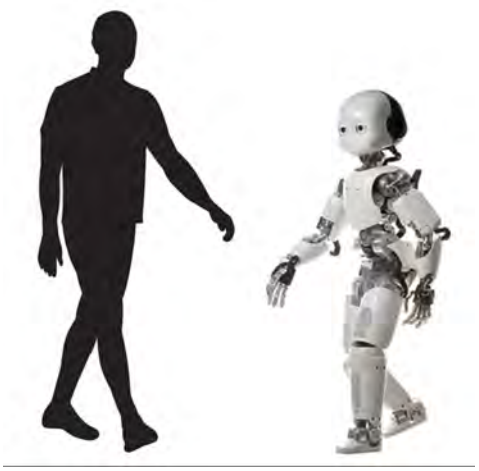

Figure 2.8 **Embodiment, graphic representation of the first pillar.**. The body has a crucial role in interaction also in HRI: a connection exists between the observer's and the observed agent's body.

A second factor shows the difference between the perception of the environment and others. In both cases, our senses receive numerous impressions, but the observer's body participates in perception differently. A deep connection exists between the body of others and the one of the observer: a connection which is clearly shown by neurobiological evidence of the mirror neuron system [183, 182]. Others' movements, actions, and even bodily reactions are received with sight, hearing, and touch but experienced by passing through the observer's body schema, which reflects and interprets others' bodily experiences.

2.5.1.2 Intentionality

Intentionality is the property of internal states for their being directed toward an object (see Figure 2.9). It is the aspect of *aboutness* that acts such as perceptions, intentions, beliefs, desires, imaginations all have: that of being perceptions, beliefs, desires, imaginations *of something*. From a phenomenological perspective [97], it is worth noticing that the intentional aspect of perception implies that the object the observer perceives is not the mental image of the object present in the world, but the real object, although perceived in a defined modality: i.e., from a specific perspective, through a unique body, based on a personal previous experience, with a certain affective state, etc. This entails

1. that when interacting with others, the observer can refer to the same environment the others perceive, although from a different perspective;

2. that the observer can understand and integrate the way others relate to the environment;

Figure 2.9 **Intentionality, graphic representation of the second pillar.** The fundamental attribute of intentionality is the *reference* of the agent toward the external world.

3. that in reciprocal interactions and shared perceptions the observer and the others can create a common ground and jointly act upon the same environment.

Tomasello described the latter as shared intentionality: “the ability to participate with others in collaborative activities with shared goals and intentions” [229]. Only by having a shared goal that keeps the same aboutness, i.e., directed toward the same object, engaging in joint actions becomes possible.

2.5.1.3 Human-Awareness

Human awareness can be termed as the consideration the robot needs to have for humans while interacting with them. This implies that, whatever role it holds, whether observer or other, the interactive abilities of the robot should be designed according to human interactive skills (see Figure 2.10). Hence, in the case of a robot "observer", it is crucial to endow the robot with perceptual abilities to understand humans by keeping track of non-verbal behaviors, inferring emotional states, and recognizing their gestures. Conversely, in the case of a robot in the role of "others", human awareness involves rather the consideration the robot should exhibit in expressing in a human-like, clear fashion its internal states. The robot’s behavior should be understandable by human eyes, and its inner states comprehensible so that humans might easily achieve Shared Perception with robots as partners.

One of the requirements for Shared Perception to emerge is the development of human-aware robots: robots capable of reading human covert inner states and expressing their own covert inner state. These abilities would humanize the interaction between humans and

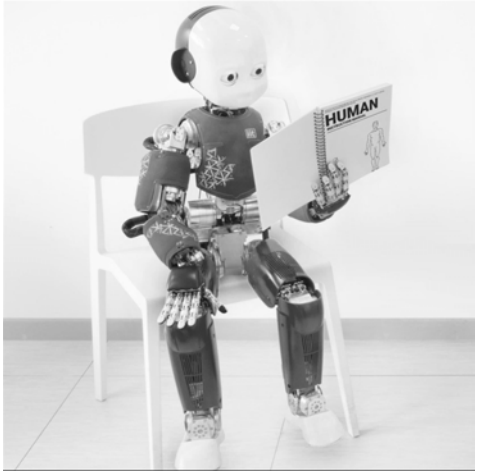

Figure 2.10 **Human Awareness, graphic representation of the third pillar.** The basis for having a human-aware robot is informing the robot with models of human behavior and mind.

robots, making it more natural and enhancing its fulfillment [196]. Therefore, although in human-robot interaction the term agent awareness may be seen as more inclusive, speaking about human awareness might be more correct because these skills are inspired by human cognition and interaction and because it is the robot that needs to be humanized instead of the human being robotized. Only if robots are developed to be aware of humans it will be possible to foster natural HRI and make robots truly human-friendly, cooperative, and assistive tools.

2.5.1.4 Hermeneutics

Hermeneutics can be conceived as the art of understanding, realized via interpretation [69]. In that, perception is a hermeneutical process because what we received from the senses is experienced only as already interpreted. A pure sensory datum is never given to experience, but necessarily appears situated in a context. The description of the perceptual process (see Section 2.3) and the influence of sociality (see Section 2.4) shine a light on how interpretation might occur in the Shared Perception process. Sensations are experienced insofar as they are perceived as unified, contextualized on previous sensory history, transformed by predictions, colored by affective states, interpreted based on others' attention, perspectives, actions, or inner states.

There is a wide range of active contextual information. For the sake of simplicity, the interpretative process within Shared Perception can be summarized in three general contexts.

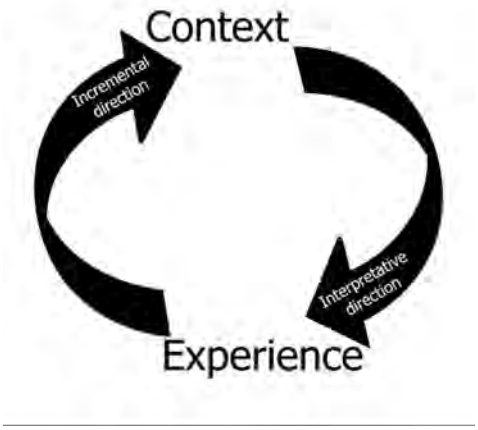

Figure 2.11 **Hermeneutics, graphic representation of the fourth pillar.** The circle between the interpretative direction and the incremental direction forms the human hermeneutical approach to the world.

1. Environmental properties, including those related to time, space, and objects situated in it.
2. Social properties connected to others, including personality, appearance, behavioral coherence, social role, and the number of agents.
3. Self-properties, including previous perceptual history, affective states, and predictions.

Another element pointed out by hermeneutics is the circular nature of interpretation. If each sensation is experienced based on a context, it is also true that contexts are generated by accumulating and associating experiences. Contexts allow us to interpret sensations and experience them. Conversely, they rise from experiences. This interdependent relation can be figured as a circle between context and experience. Taken into our perspective, the repetition of similar events turns into the uniformity of a context that can therefore be known and restricted from experience. In addition, multiple perceptions can be gathered and grasped just because they stand in the background of the same interpretative context. It becomes clear that perception requires both elements: a context to interpret single phenomena (interpretative direction) and the process of learning from experiences to develop a deeper understanding of contexts or the comprehension of novel ones (incremental direction) (see Figure 2.11).

2.5.1.5 Sharedness

The idea of Sharedness, which forms the basis of Shared Perception, may have different connotations and nuances of meaning. Taking inspiration from previous literature that

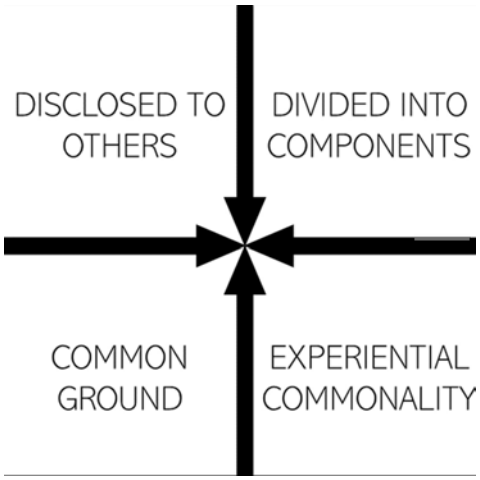

Figure 2.12 **Sharedness, graphic representation of the fifth pillar.** Multiple meanings of the word "shared" are at the basis of Shared Perception.

deepened this topic [32, 56], the present framework is grounded on four different senses of the term 'shared' (see Figure 2.12).

1. The first meaning concerns the idea of disclosure and communication. From this perspective, something shared needs to be actively, albeit not necessarily consciously, *communicated* by the others to the observer.

2. The second connotation refers to something being *partitioned*, i.e., something that can be divided into different components, each belonging to someone. In this sense, the parties (observer and others) share something because each keeps an element of it.

3. The third nuance expresses the agreement and consensus reached by all parties over what is shared. In this sense, what is shared is the *common ground* whereon the parties can communicate and interact.

4. A fourth meaning implies the emotional sphere of experiencing and feeling the *commonality of inner states with others*. From this perspective, all the parties are aware of sharing inner states with others and, accordingly, a closer connection and influence among them is established.

Although the concept of Sharedness comprises different shades, the interactive process in which Shared Perception takes place often allows their integration. By disclosing inner states through their behavior (1st meaning), the parties achieve a common ground on the environment (3rd meaning). Concurrently, a full common ground is never given. Hence the

presence of different agents with different points of view represents the intersubjective perception of the environment, partitioned into different perspectives (2nd meaning). Eventually, the commonalities achieved through Shared Perception may be experienced by the parties and have a positive impact on their engagement (4th meaning).

2.5.2 The dynamic composition of Shared Perception: towards a triangulation

The factors outlined thus far apply to describe the process of Shared Perception. The contextualization in socio-cognitive HRI defines the actors of the frameworks: humans and social robots. Although Shared Perception does not ideally limit the number of actors, for the sake of simplicity it is possible to reduce the field to one "observer" and one "other": roles that can be interchangeably interpreted by the human or the robot. Still, to compose the structure of Shared Perception, a third element is required: the environment. Besides the above-mentioned meanings of "sharedness", here we encounter yet another crucial connotation of this word underlying all the others. A "shared" environment is the ground zero of Shared Perception: the environment that the observer and the other share. In this sense, "sharing" does not imply, but enables the interaction among agents. As the third element of Shared Perception, the environment generically represents the place where the two agents are, but the term can be used to represent any third element that is present or happens in the environment, might it be an object, an event, or a third agent.

Shared Perception is therefore founded on three elements, The observer, the other, and the environment, each playing a different role. Respectively, the subject, the co-subject, and the object of Shared Perception. From a dynamic point of view, Shared Perception is thus composed (see Figure 2.13 for illustration). While the observer perceives the environment, it concurrently perceives the other. The embodied nature of the two agents and the consequent mind attribution to the other drive the observer to get the intentional relation of the other toward the environment and integrate it with its own intentional relation toward it. In this way, the other starts from being perceived by the observer and turns out to be a co-subject of Shared Perception or, in other terms, the observer's perception of the environment results to be triangulated with the other.

Shared Perception can be therefore seen as a triangled integration of information: from the world (environment), the self (observer), and other agents (other). The description of the perceptual process (see Section 2.3) shines a light on how we integrate elements related to the objects and the self. In addition, others can represent another source of information, as

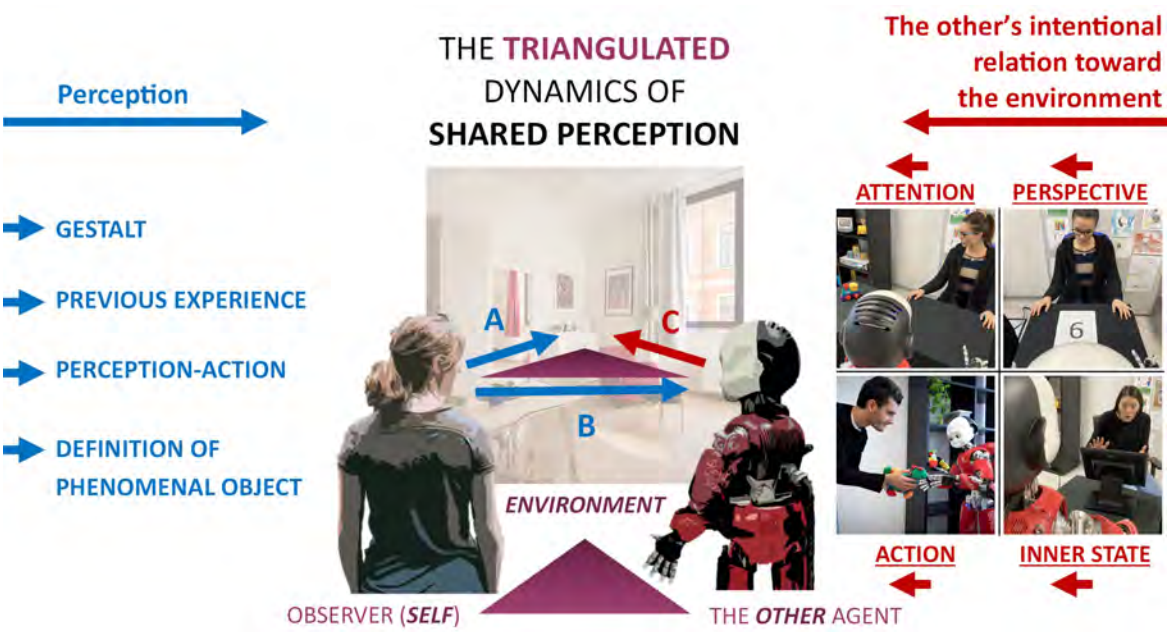

Figure 2.13 **Illustration of the Shared Perception dynamics**. Shared Perception can be described as the observer's perception that emerges from the triangulation among three sources of information: the environment, the self, and the other. As a first step, the observer perceives the environment (arrow A in blue): multiple elements (some of them listed on the left in blue, as described in Section 2.3) take part in such a process, which is not only a passive reception of stimuli because features referred to the object (environment) and the observer (self) are integrated together. If another social agent is present in the same environment and perceived by the observer (arrow B in blue), the other's intentional relation toward the environment (arrow C in red) is obtained by the observer and integrated into the whole perceptual process. As described in Section 2.4, the others affect the observer's perception through their attention, perspective, actions, or inner states (listed on the right in red). This way, the triangulation among the observer, the other, and the environment takes place. Either the human or the robot could play the role of observer. The blue arrows (A and B) that represent the observer's perception do not move from the perceived object toward the observer but vice versa. In this way, the active dimension of perception is underlined because perceiving means actively contributing to the creation of the perceived object.

illustrated by the four components of others' behavior that we integrate into the perceptual process (see Section 2.4).

The idea of triangulation is inspired by the philosopher D. Davidson [46], according to whom three varieties of knowledge are interdependent: knowledge of the world, of one's mind, and of others' minds. Two principles rule the knowledge of other minds: the Principle of Coherence and the Principle of Correspondence. For the former, the observer tends to attribute logical consistency to the other's mind. For the latter, it tends to assume that the

other reacts to the world likewise itself. Hence, the knowledge of others appears strictly intertwined with the knowledge of the world. It requires a constant comparison with the world shared with them. Secondly, as it becomes clearer when considering child development, knowledge of the cause and the content of our thoughts becomes possible only after sharing the reaction to external stimuli with others. The knowledge of others is intertwined with the knowledge of one's mind as well. Knowledge about the propositional content of our thoughts is not possible without the other two forms because propositional thought is made possible by communication and reference to the external world. Alongside, the knowledge of others requires the knowledge of one's propositional thought because one can attribute thought to others only by matching their behavior with its own propositional thought.

Inspired by the interdependency of these three varieties of knowledge, the Shared Perception dynamics can be expressed along the same core directions: the self, the environment, and the other. As a result, through this triangulation Shared Perception produces three effects in which all these three terms are involved.

1. Self. Subjective awareness of the environment is enhanced by others. Integrating at least one of the four components (attention, perspective, action, mental states) of others into the perception of the environment produces in the observer an augmented knowledge of the environment in terms of features, objects, elements, and aspects not yet discovered or understood. In this sense, a shared (communicated) perception produces an augmented understanding of the environment.

2. Environment. The environment is perceived as common ground between the self and the other. The integration and the comparison between the observer's and the other's intentional relation toward the environment allow the two interactants to perceive the environment from a common viewpoint and, therefore, to meaningfully communicate and effectively collaborate. A shared (common) perception produces a successful interaction.

3. Other. Understanding others is adjusted and enhanced based on previous experience and environmental conditions. Following the Principle of Coherence, previous experience with one person can be used by the observer to understand their behavior in a subsequent interaction. Following the Principle of Correspondence, the observer's personal experience with the environment can be used to assume the other's experience toward it. Taken together, if the other enhances the observer's perception of the environment, it is also true that the observer understands the other through its own previous experience and through the other's intentional relation toward the environment.

Interacting with the shared (same) environment and having shared (common) feelings toward it produces a deeper comprehension of others.

2.6 Conclusion

It appears that Shared Perception is a core component of human relations but, if we aim for robots to be able to interact naturally with humans, its role should be equally important also in HRI.

From the perspective of the robot's observer, the interpretation of humans' behavior and the understanding of their intentional relation toward the environment may represent a possible way to enhance the robot's awareness of the environment. As well as for humans, the ability to integrate others into perception entails several benefits. Perception is augmented because objects not yet detected, aspects not yet understood, or events not properly interpreted may become so. Moreover, if perceiving others improves perceiving the environment, then also the opposite direction is true for the triangulation principle. A better perception of the environment, as a retroactive effect, may produce a deeper understanding of others. Moreover, thanks to this process a common ground emerges between the observer and other agents, which is the only way for meaningful and effective interaction.

The same benefits may result from the perspective of a human observer because the process of Shared Perception is allowed and elicited by the robot's body, likewise the humans'. Thanks to its body, the robot is able to express its intentional relation toward the environment, and the human observer can understand it: the process of Shared Perception can start.

The present framework presented Shared Perception, showed its crucial role in human-Robot interaction, outlined the elements needed for its emergence in humans and for its development in robots, described the mechanisms of the process, listed the factors and components that need to be integrated, and eventually, exposed its effects on the interaction. In the context of Human-robot interaction, this framework could guide experimental research to investigate humans' perception while interacting with robots and, from an opposite perspective, direct the development of robots' socio-perceptual skills inspired by Shared Perception.

Chapter 3

Shared Perception and Context Dependency: a user study to investigate the impact of a social robot on human visual perception of space

3.1 Introduction

Human perception integrates sensory information and predictions about the external world, a phenomenon that Helmholtz described in terms of unconscious inference [92]. Thus, sensory inputs are influenced by the previous experience organized along internal models acting as priors. A large body of research established that these two sources of information are integrated through Bayesian principles in many tasks, such as perception of an object (for a review, see [117]), visual speed [246, 214], time intervals [109, 38, 113, 184], categories [98], lengths [195], and spatial localization [18]. The use of priors improves the reliability of perception, reducing the overall noise, and is often considered to reflect a statistically optimal computation [202]. The influence of priors increases in the presence of low reliability of sensory input to cope with the uncertainty of the external world. For instance, this happens when the noise is due to an increased vagueness of the sensory information [18] or in people with lower perceptual acuity [38].

The phenomenon of Context Dependency, which had already been described by [93] with the name of Central Tendency, has been modeled in terms of Bayesian prior integration [109, 38, 195, 113, 184]. When exposed to a series of stimuli of the same type, the perception

of one stimulus is affected by the stimuli perceived before, so their reproduction tends to gravitate toward their arithmetic mean. Therefore, perception is affected by the previous experience, built throughout the exposition of the entire series of stimuli. Such experience acts as an internal predictive model, a prior, on the incoming stimuli to reduce the variability of responses. Albeit at the expense of accuracy, prior influence produces an increased precision, resulting in a minor perceptual error as an overall effect. Nonetheless, this beneficial effect on individual perception could hinder the efficacy of an interaction. Relying on previous personal experience could cause misalignment with another agent having a different prior history, preventing, for instance, effective coordination.

Social interactions require establishing a common ground with the partner [225]. Without it, interactants would make nonsense of any verbal or non-verbal communication, causing misunderstandings, ambiguities, lack of coordination, or perceptual mistakes. Even though different people might experience different perceptions of the same environment – opposite perspectives or the most varied emotional states – they commonly succeed in interacting with others by bridging these differences. How is this achieved when the difference between two individuals' perceptions stems from different prior histories?

In this study[1], we address the question of the role of internal predictive models on perception in a shared environment. Do humans maximize individual perceptual stability using internal priors, or do they align perception with the partner to facilitate coordination by limiting the reliance on individual priors?

Social interactions shape several human perceptual and cognitive processes. From first months of life, selective attention is influenced by the direction of the partner's gaze [14]. This is the basis of an ontogenetic process that will lead to other interactive behaviors [229]. For instance, the ability to take the perspective of another person [191, 88, 13] seems to have its origins in this developmental process [149]. Furthermore, sociality impacts gaze movements [178], memory processes and information encoding at different levels [204, 205, 178, 57]. It affects the processes of perception-action underlying joint-action (e.g. the Joint Simon Effect [120, 199]), and influences the perception of space [150, 151]. Therefore, we believe that a social context could significantly shape also basic perceptual mechanisms, such as Context Dependency.

To address this kind of question, which explores the concept of Shared Perception, it is necessary to move the investigation from an individual, passive approach to an interactive shared context. To this aim, we propose to employ a humanoid robot as an experimental tool to investigate how perceptual mechanisms change during social interaction. Cognitive

[1]The outcomes of this work have been published in [141, 142].

science research studying the influence of a social context on perception may benefit from the use of embodied artificial agents such as robots [198]. Such complex sensory-motor devices allow for generating controlled and precise actions in a repeatable manner. That enables the experimenter to replicate the rigorous control of stimuli traditionally adopted in the standard perceptual investigations within an interactive setting. This approach grants a degree of reproducibility of the (social and non-social) stimuli, which human actors cannot guarantee. Robots ensure an ecological layout to experimental settings thanks to their embodied presence in the shared physical space, instead of the virtual presence of an agent shown on a screen.

Extensive evidence shows the feasibility of the approach, demonstrating that robotic platforms can evoke social effects on humans, similar to those observed in human-human interactions. For instance, a robot can establish joint attention with users and elicit inferences about the intended referent [209]. Its behavior induces the same brain processes as if it was provoked by a human agent [123]. It has also been shown that robots elicit the same cognitive mechanisms of visual perspective-taking (VPT) that usually are elicited by human agents. The human partner spontaneously takes the visual perspective of the robot on a shared target, primarily when the robot directs its gaze or performs a reaching action toward it [258]. Moreover, Joint Simon Effect has also been found during interaction with robots [212, 216], suggesting that humans implicitly represent robots' actions as other humans' ones during joint actions. For other effects induced by robots, similar to those elicited by humans see Section 2.4.

Starting from this perspective, the present user study investigated the impact of social interaction on the perceptual processes of prior integration. Participants have been asked to perform a perceptual task – estimating the length of a stimulus – in a social and non-social scenario. The study aimed to assess whether human perceptual performances change and whether participants followed the prediction of a Bayesian model of Context Dependency. To achieve this, we employed a humanoid robot as a stimulus demonstrator to keep the same stimulation and just manipulate the context making it either social or non-social.

3.2 Methods

The present research was conducted to evaluate if space perception changes when the perceptual task is not performed in isolation but with another agent. More specifically, the objective consisted of understanding whether, during social interaction, human perception complies with the same principles of optimization it follows in individual scenarios [38, 195]. To this aim, we designed a user study to explore how the perceptual phenomenon of Context

Dependency is affected by interaction with a humanoid robot acting as a mechanical or social agent, depending on the experimental session.

3.2.1 Participants' demographics and Ethics

The experiment involved 30 participants (15 F, 15 M) over the age range of 19-46 years (M=28, SD=6). 37% of them had already been exposed to interaction with the robot employed for the research (iCub). Nobody was aware of the purpose of the study. Due to technical problems, 3 participants could not finish the experiment, whereas other 2 participants had been excluded as outliers (see Section 3.2.6.3) so that in the end the sample was composed of 25 participants (13 F, 12 M). All of them signed a written informed consent before the experiment and received an honorarium previously agreed of 15€ for their time. The research had been approved by the regional ethical committee (Comitato Etico Regione Liguria).

3.2.2 Procedure

The study consisted of a reproduction task and involved three counterbalanced within-subject conditions. In two of them, the participants interacted with the robot, whereas another one was performed individually. The experiment lasted approximately 90 minutes. Participants had been previously informed of the duration of the experiment.

3.2.2.1 Experimental setting

Participants performed all the tasks in the same experimental room where they sat at 50 cm from a touchscreen placed on a base 75 cm tall. During the task performed with the robot, the robot was placed on a fixed platform at 20 cm on the other side of the touchscreen, whereas during the individual session, it was hidden behind a curtain. Figure 3.1.A-C reports a schema of the experimental room. Another curtain hid the experimenter's station with a table and the computers connected to the touchscreen and to the robot. Blinds were closed and the room was lit up with artificial light in order to ensure the same lighting conditions for all participants.

In this study, we assessed how prior influence is altered when perceiving stimuli provided by another agent. To this aim, we needed an agent who acted as stimuli demonstrator reliably and consistently with all participants. We thereby employed the humanoid robot iCub [147], which is capable both to show a social behavior and to generate controlled and precise actions to replicate the rigorous protocol adopted in standard perceptual studies. The behavior of

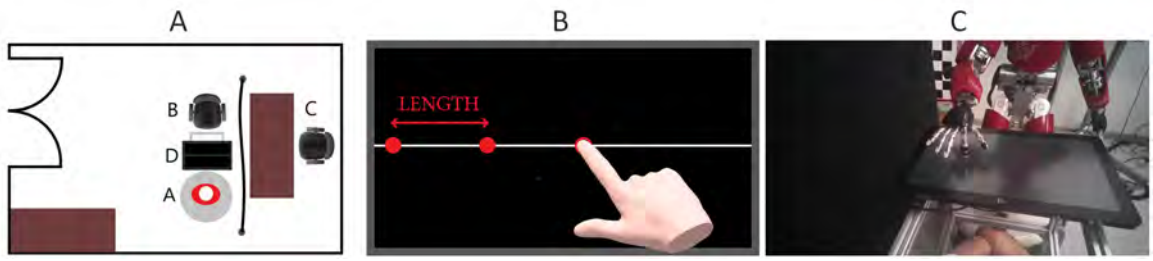

Figure 3.1 **Experimental setting for reproduction task.** Figure A: Setting of experimental room: (A) iCub robot's place, (B) participant's place, (C) experimenter's desk, (D) Touchscreen. Figure B: Description of Individual length reproduction task. Two dots are presented consecutively on a white line on a touchscreen, showing a certain length. Participants had to keep the second dot as a reference and to touch the screen in a third point, to reproduce the length of the stimulus. Figure C: iCub from participants' perspective while touching the screen to present stimuli: images were obtained from Tobii Pro Glasses 2 recording.

iCub was controlled to perform humanlike minimum jerk movements with an average hand speed of about 0.1 m/s. Specifically, the robot iCub presented the stimuli to participants by touching the screen and moving its torso and right arm according to models of biological motion.

A widescreen LCD Touchscreen Monitor ELO 2002L 20-inch was employed with a resolution of 1920x1080 px for an active area of 436.9mm x 240.7mm, at a frequency of 60Hz and Response Time of 0.02 sec. The monitor was positioned horizontally: it showed the stimuli to participants and recorded both the touches of the robot and the responses of participants. It was programmed with MATLAB 2019a with Psychophysics Toolbox Version 3 (PTB-3) and controlled by a Windows 10 pc. To record participants' gaze information during the interaction with iCub, we asked them to wear a Tobii Pro Glasses 2 (100 Hz gaze sampling frequency).

3.2.3 Experimental Sessions

To test the experimental hypotheses, we set up different sessions. An individual task of length reproduction served as a baseline to assess participants' level of Context Dependency. The other two sessions were performed with the robot acting differently, as a mechanical or social agent, to determine how social interaction affects Context Dependency in perception.

3.2.3.1 Individual length reproduction task

In the individual length reproduction session, the participant's task was to reproduce the lengths indicated by two dots presented on the screen by touching the screen on a third point (see Figure3.1.B). Specifically, the reproduced distance between the second dot and the point touched by participants – should be equal to the presented length. The stimuli were presented as two consecutive red circles of 1 cm diameter lasting 0.6 s each and appearing with an interval of 2 s. The first dot was presented at a variable distance from the left border of the screen (0.5–3.5 cm, randomly selected). The second dot was shown at the right of the first one, at a distance of 11 different lengths from 6 cm to 14 cm with a difference of 0.8 cm each. Each distance was presented 6 times, randomly, for a total of 66 trials with additional 3 practice trials. After the response, another equal red disk appeared at the touched point, but no feedback was provided about the accuracy of the response.

3.2.3.2 Length reproduction tasks with the robot

In the two main sessions of the experiment, participants interacted with the humanoid robot iCub. iCub acted as a stimulus demonstrator touching the screen in the two endpoints of the lengths (see Figure 3.1.C). Participants' task was the same as in the individual length reproduction task (see Figure 3.1.B). The touchscreen did not show any light in the points where iCub or the participants touched. The robot was programmed to present the same stimuli as in the individual task. Whereas participants' task was the same in both conditions, the behavior of the robot changed from one condition to the other. Indeed, for a correct evaluation of the impact of sociality on perception, it was necessary to compare two conditions where the very same sensory inputs were presented as stimuli, and only the nature of the presenter (social vs mechanical) was manipulated. In this case, the stimuli were always provided by the robot's finger indicating two points on the touchscreen, with the very same kinematics in two different conditions, "Social" and "Mechanical".

3.2.3.3 Individual Length Discrimination for perceptual ability check

Beyond the three reproduction sessions, an additional length discrimination task aimed to test the perceptual acuity of the participants in order to find possible outliers. Three red disks of 1 cm diameter appeared for 0.4 s in sequence with an interval of 1.5 s on a white straight line crossing the screen at its central height. After stimulus disappearance, subjects had to judge whether the longest segment was the first, delimited by the first and the second disk, or the second one, delimited by the second and the third disk, by typing respectively "1" or

"2" on a keyboard located between them and the touchscreen. Participants performed this task for 66 trials. One of the two distances (standard) always measured 10 cm, while the other (comparison) changed from trial to trial according to a QUEST adaptive procedure [245]. This design represents a very simple measure of length discrimination, where priors do not influence performance. The proportion of times in which the comparison interval was judged longer than the standard was plotted as a function of comparison amplitude and fit by a cumulative Gaussian distribution. The standard deviation of the fitted Gaussian represents the perceptual threshold, which is the minimal difference between two lengths that the participant can reliably distinguish.

3.2.4 Characterization of robot's behavior for interaction design

Since the research aimed to study the perceptual alteration induced by sociality with the aid of a robotic stimuli demonstrator, we decided to differentiate as much as possible the way participants perceived the robot in the two conditions. Implicit behavioral and verbal cues of the robot were therefore combined with explicit priming of participants about the robot's intentionality and skills.

In the social condition, iCub acted as an interactive social agent[2]. Its left eye camera was turned on to track participants' faces and establish mutual gaze before starting the task, after its end, and between one trial and another, to give an implicit idea of turn-taking (see Figure 3.2). To enhance the impression of animacy, for the entire duration of the interactive condition, the eyelids were blinking. Moreover, iCub showed emotions with its facial LEDs: it mostly smiled with a friendly expression, unless while touching the screen, when it was programmed to appear focused on the task. Through the iKinGazeCtrl Module [187], iCub also exhibited natural oculomotor coordination with its hand by directing its gaze in advance toward the point it was going to touch. Before starting the practice trials, iCub welcomed participants and explained to them the task (in italian): *"Hi, I'm iCub! Now, we will play together. I will touch the screen twice, and you will touch the screen a third time to replicate the distance. Are you ready?"*. Then after 1/3 and 2/3 of trials, the robot incited participants with these words: *"Well done! Keep it up!" and "Come on, there are only a few more trials left, keep focused"*. Finally, at the end of the task: *"Thank you for having played with me! It has taken a bit of a long time, but you are helping us a lot! See you next time"*. During the speech, the mouth-LEDs simulated the lips movement in coordination with the words iCub was saying.

[2] I wish to thank Alexander Mois Aroyo for his help in programming the iCub's behaviors

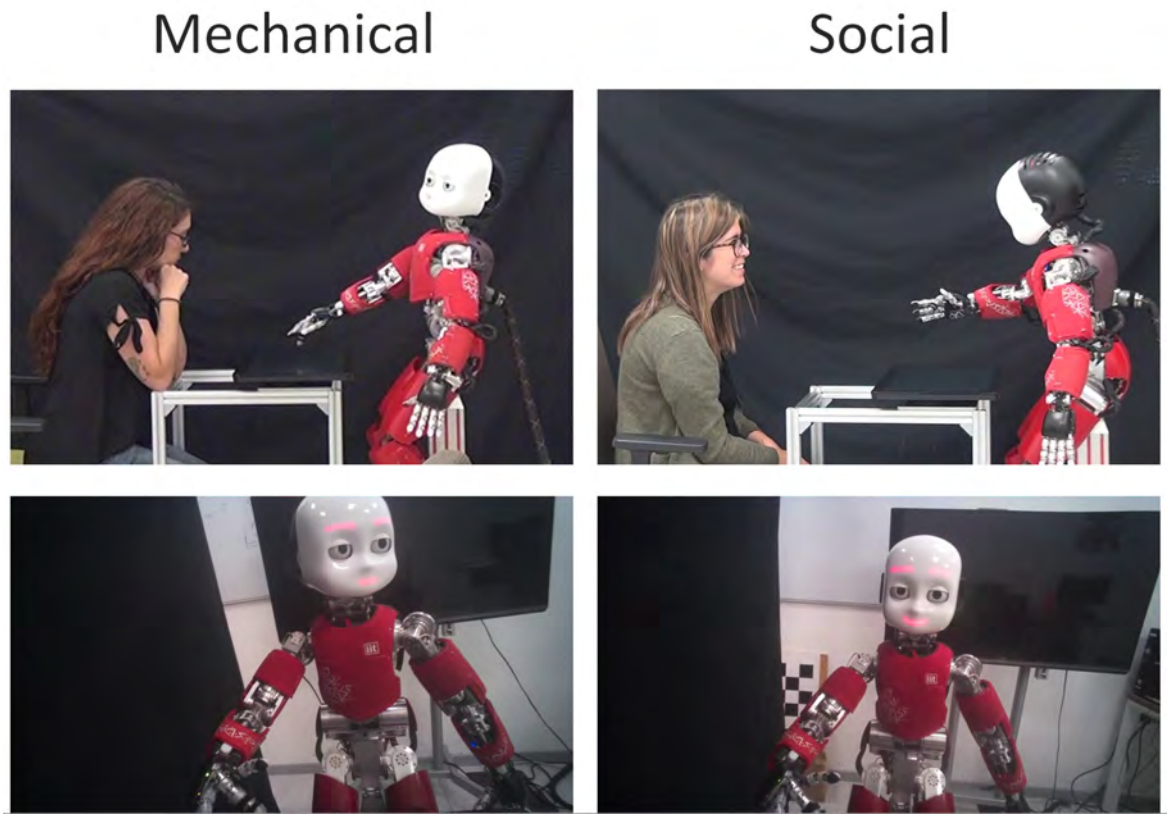

Figure 3.2 **The interaction with iCub during the reproduction task**. On the left, pictures of the robot behaving mechanically taken by an external camera (above) and by the Tobii Pro Glasses eye tracker that participants wore during the task (bottom). On the right the same pictures with the iCub behaving socially. The head direction of the mechanical robot was fixed and turned away from the participants, whereas the social robot could look at the screen and exchange mutual gaze with participants.

Conversely, in the mechanical condition, iCub acted as a mechanical agent without showing any social features. To this aim, iCub head joints were fixed so that its head was turned away from the participants (see Figure 3.2). This behavior was designed to show that the robot had no awareness of the environment or the task. Also, face-LEDs were static, so the robot appeared without emotions, and the robot did not talk. The only parts that were moved were the joints of the torso and the right arm, like a robotic arm. To strengthen the differentiation of the two conditions, the experimenter diversified the introductory explanation of the task when talking about the robot. Outside the experimental room, in the social condition, the researcher introduced the session in this way: *"Now iCub is fully working, with its social intelligence on. Its cameras are switched on to look at you and the screen. It will be showing you two positions on the touch screen. Please*

reproduce the distance between these two points by pressing the touchscreen in a third one at an equal distance from the last shown by the robot". Conversely, before starting the task, participants were instructed with these words by the experimenter: *"In this session, iCub's social intelligence is turned off. The computer is just driving its hand motions in a predefined pattern. It will be touching two positions on the touch screen. Please, reproduce the distance between these two points by pressing the touchscreen in a third one at an equal distance from the last one"*.

3.2.5 Questionnaires

We collected data from a set of questionnaires through Google Forms. The first questionnaire was compiled before coming in the laboratory and included some questions about participants' previous experience with robots, the Italian version of TIPI test on participants' personality [36], the Autism-spectrum Quotient test (AQ test) [15, 189] to measure the degree to which adults with normal intelligence have the traits associated with the autistic spectrum, and the NARS questionnaire, to evaluate the attitudes of participants towards robots [221].

Another set of questionnaires was submitted after each session with the robot to check the manipulation effect of the robot's behavior and explicit priming. To this aim, once participants ended each task with the robot, they were asked to go out of the room and fill a form of questions online. On this occasion, we delivered the Inclusion of Other in Self-scale (IOS) questionnaire [8] to assess how close to iCub participants felt during the task; the Godspeed questionnaires with the sub-scales Anthropomorphism, Animacy, Likeability, and Perceived Intelligence [17] and the subscales Mind experience and Mind agency of a Mind perception test [87, 60]. We proposed all of them on a 7-points Likert scale.

At the end of the experiment, a final questionnaire for debriefing was provided to participants to collect their opinions and feedback about the tasks and the behavior of iCub.

3.2.6 Data Analysis

3.2.6.1 Length Reproduction

To investigate the phenomenon of Context Dependency, we analyzed the reproduced lengths following a well-established approach [38, 195]. The influence of prior experience on sensory stimuli, which occurs as the integration of different kinds of information, can also be interpreted as the dependence of perception on its context. For instance, in visual perception of space, perception of a visual stimulus is affected by distances experienced before, which

cause a perceptual bias. The overall effect resulting from such integration is thereby a regression of all perceived stimuli toward the mean of the presented stimuli, which act as prior built during the exposition to all the set of stimuli. In this way, the long distances are perceived as shorter than they are and vice versa. Regression Index is a direct measure of the degree of Context Dependency: it is computed as the difference in slope between the identity line (stimuli-correct responses) and the best linear fit of the reproduced values plotted against the related stimulus (see Figure 3.3). The index varies from 0 (no regression) to 1 (complete regression). Specifically, in our study, the stimuli were presented to participants so that their arithmetic mean was 10 cm.

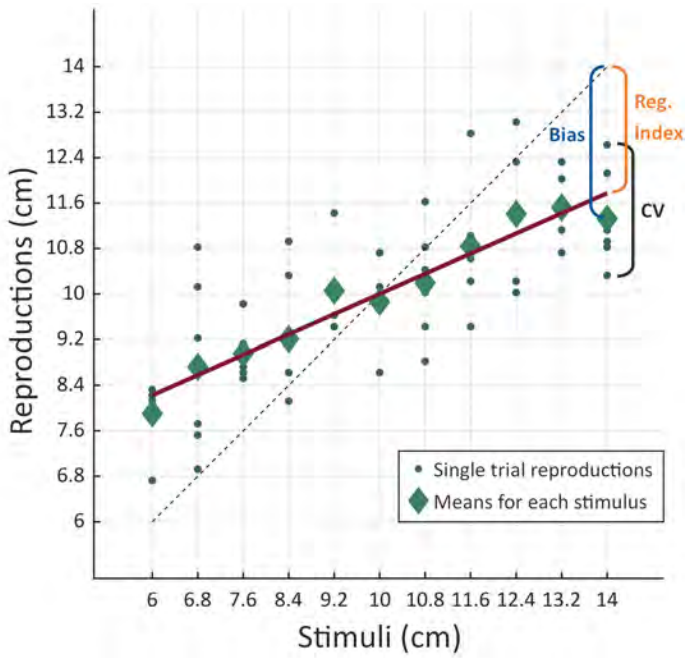

Figure 3.3 **Illustrative plot of the data of a length reproduction task.** Reproduced lengths are plotted against the related stimuli. The regression index is calculated as the difference between the slope of the linear fit of the ideal reproductions (identity line) and the slope of the linear fit of the real data. For each stimulus, we also measured the average bias and the coefficient of variation of the related responses.

The present research also aimed to assess and model the perceptual errors associated with the phenomenon of Context Dependency, hence we portioned the total error of responses into two parts: the bias and the coefficient of variation (CV) that respectively measured participants' accuracy and precision. First, a constant bias was removed from each n-th response of each i-th stimulus ($R_{i,n}$) by subtracting the average response of all trials (R_c) and summing the length of the average stimulus ($\bar{S}$).

$$R_{i,n}{}' = R_{i,n} - R_c + \bar{S} \tag{3.1}$$

Then, for each i-th stimulus, we measured bias as the difference between the average response for that stimulus (R_{Mi}) and the stimulus (S_M), in absolute value, normalized for the average stimulus of the entire session ($\bar{S}$). In the robot sessions, since motor noise caused a slight imprecision in the stimuli demonstration, we used the average stimulus presented by iCub for each of the 11 lengths (S_{Mi}).

$$BIAS_i = \frac{|R_{Mi} - S_{Mi}|}{\bar{S}} \tag{3.2}$$

The CV of responses to each stimulus was calculated from the standard deviation of the responses to that stimulus, again normalized for the average stimulus of the entire session $\bar{S}$.

$$CV_i = \frac{\frac{\sqrt{\sum (R_i{}' - \bar{R}_i{}')^2}}{N}}{\bar{S}} \tag{3.3}$$

Finally, the normalized total error is calculated for each stimulus as the root-mean-square error (RMSE) from the bias and the CV

$$RMSE_i = \sqrt{BIAS_i^2 + CV_i^2} \tag{3.4}$$

Statistical analyses of the data related to perceptual errors in the three conditions were conducted using the Linear Mixed Models in R with the following libraries [126, 135].

3.2.6.2 Gaze analysis

To assess possible variations in the way participants visually interacted with the robot in the two conditions, we analyzed data of participants' gaze gathered through a gaze-tracker, the Tobii Pro Glasses 2, during the task performed with iCub. This information also served as an additional behavioral check of the manipulation of iCub's social features to understand whether the robot was also recognized implicitly by participants. The number of times participants looked at iCub's face during the experiment served as a measure of participants' involvement during the interaction. To extract these data, we first obtained the images of the iCub's face from Tobii recordings. Then, we trained the software of the gaze-tracker (Tobii Pro Lab) to recognize the face of iCub as a region of interest in the recordings to check whether participants' looks stopped on the robot's face (see Figure 3.6.B). This way, we counted the percentage of times in which participants looked at iCub's face during each

session. We assessed such a percentage by counting the number of trials in which iCub's face was looked at at least once. Specifically, the measure was taken for two kinds of time intervals: the interval between the first and the second touch of iCub (during trials) and the interval between the second touch of iCub and the first one of the subsequent trials (between trials).

We conducted statistical analyses to compare participants' gaze data in the two tasks with the robot with Jamovi 1.6.1 [106]. Data have been extracted using Tobii Pro Lab Software and Python with Pandas Data Analysis Library. Due to technical problems with the device and because some of the participants wore their glasses, we could analyze only 15 participants from our sample.

3.2.6.3 Perceptual ability check and outliers

We organized a perceptual task of length discrimination to assess whether participants were able to perceive the visual stimuli reliably or whether all their performances should be discarded (See Section 3.2.3.3). Specifically, we decided not to analyze participants who revealed not being able to discriminate a distance smaller than 4 cm, which was the difference between the mean stimulus of the reproduction task and the extreme ones.

We also decided to exclude participants whose performance in the reproduction tasks exceeded the average performance of all participants of at least 2.5 times the SD of the sample. We removed two participants from the sample after this last screening.

3.2.7 Bayesian Modeling

Context Dependency is a perceptual phenomenon that can be explained as the integration between sensory information (each current stimulus) and priors (built on the stimuli already perceived). Previous research demonstrated that such phenomenon can be described in a Bayesian fashion and follow Bayesian principles of optimality [109, 38, 195, 113]. Specifically, although leading to inferior accuracy in the outcome of the perceptual process, the influence of priors enhances precision – and the overall total error – by reducing the variability of the responses.

The present research aims therefore to analyze the influence of priors on visual perception of space by connecting with previous studies and, for the first time, to assess the effect of sociality on Context Dependency with a Bayesian approach.

In this perspective, following the approach proposed by [38], the perceived length of a stimulus (Posterior) can be modeled as a Gaussian defined by μ_R and σ_R, and resulting from

the product of other two Gaussians (see Figure 3.4): 1) the current noisy sensation of the stimulus length, represented by the Likelihood, and 2) the Prior, which is an estimate of the series of stimuli previously perceived.

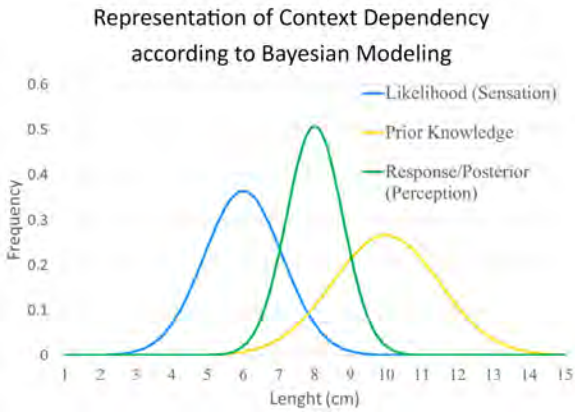

Figure 3.4 **Representation of Bayesian Model.** (Modified by [38, 195].) Perception (Posterior distribution) is described as a Gaussian resulting from the integration between the Likelihood distribution of the stimulus of length μ_L with sensory precision of σ_L and the Prior distribution centered in μ_P with a weight of σ_P.

For each stimulus, the Likelihood function is modeled as a Gaussian centered on the actual length of the stimulus (μ_L) with standard deviation (σ_L) corresponding to the sensory precision of each participant. The Prior is modeled as a Gaussian distribution with the mean (μ_P), corresponding to the average stimulus of the series, and an amplitude (σ_P) that represents the weight given to the prior during perception. Thus, according to the model, given a fixed prior width, the observers' response is derived as a function of their sensory precision: the better it is (i.e., the narrower the likelihood distribution is), the nearer the response will be to the sensory information. Conversely, the worse observers' sensory estimate is, the closer their response will move towards the prior.

Given these premises, according to Bayes' rule, the mean and the standard deviation of the posterior distribution can be respectively calculated as

$$\mu_R = \mu_L - \frac{\sigma_L^2(\mu_L - \mu_P)}{\sigma_L^2 + \sigma_P^2} \tag{3.5}$$

$$\sigma_R^2 = VAR = \frac{\sigma_L^2 \sigma_P^2}{\sigma_L^2 + \sigma_P^2} \tag{3.6}$$

From Eq. 3.6 it should be noted that by construction σ_R is smaller both than σ_L and σ_P, evidencing the enhanced precision provided by the optimal integration. On the other side,

considering a series of stimuli of length S_i, the bias for a specific stimulus can be calculated using Eq. 3.2 and 3.5.

$$BIAS_{Si} = \frac{\sigma_L^2(S_i - \bar{S})}{\sigma_L^2 + \sigma_P^2} \tag{3.7}$$

Whereas, considering all the series, the bias would result as

$$BIAS = \frac{\sigma_L^2 \sqrt{\frac{\sum_i (S_i - \bar{S})^2}{N}}}{\sigma_L^2 + \sigma_P^2} \tag{3.8}$$

From Eq. 3.6 and 3.8, the total error of the observer can be therefore calculated with 3.4.

From the data obtained in the three reproduction tasks, it has been possible to model the perception of participants, compare our results with the previsions of the model, and understand how social interaction impacts on the use of prior knowledge.

The analyses and the simulation of the Bayesian Model were conducted with MATLAB 2020A.

3.3 Results

This study was founded on the primary hypothesis that interaction with a social agent plays a role in how humans perceive space. We aimed to assess whether interactive scenarios impact human integration of visual information with prior and, if it happens, how error parameters of perception, namely, accuracy and precision, are affected.

3.3.1 Manipulation check

From the questionnaires completed after each interaction with the robot, we could verify whether iCub's behaviors in the "mechanical" and "social" conditions were effectively perceived as significantly different. Table 3.1 reports all the scores of the scales provided in the questionnaires and the statistical results from Wilcoxon Signed-Rank tests. When iCub behaved socially, it was perceived as significantly more anthropomorphic, animate (see Figure 3.5), intelligent, and likable. In addition, in that condition, participants more extensively attributed to him a mind and experience. Finally, they also felt closer to him.

Behavioral measures of gaze collected through the Tobii Pro Glasses 2 confirmed that participants recognized the diverse behavior of iCub also implicitly, not only when asked through questionnaires (see Figure 3.6.A). They looked at the face of the robot significantly

Table 3.1 **Manipulation check.** Results from the questionnaires provided after each task with the robot to check whether the manipulation of the robot's behavior was correctly perceived by participants. The fourth column reports results from Wilcoxon Signed-Rank tests to compare the two conditions with the robot.

Feature	**Mechanical**	**Social**	**W. S-R. test**
Anthropomorphism (5-35)	M=12.8, SD=4.81	M=19.2, SD=5.97	Z=-3.78, p.001
Animacy (5-35)	M=13.0, SD=5.41	M=21.7, SD=5.31	Z=-4.38, p.001
Likeability (5-35)	M=22.0, SD=7.18	M=28.8, SD=21.6	Z=-3.72, p.001
Perceived Intelligence (5.35)	M=21.6, SD=4.98	M=24.6, SD=3.79	Z=-3.18, p.005
Mind experience (4-28)	M=7.04, SD=4.74	M=12.4, SD=7.66	Z=-3.55, p.001
Mind Agency (4-28)	M=11.9, SD=5.49	M=16.7, SD=6.97	Z=-3.57, p.001
Inclusion of other in the self-scale (IOS) (1-7)	M=2.84, SD=1.52	M=4.44, SD=1.42	Z=-4.09, p.001

more often in the social condition than in the mechanical one, both during trials, i.e. in the time interval between the first and the second touch of iCub (about 36% vs 8% of trials, Wilcoxon Signed-rank test: Z=120, p<0.001), and between trials, that is in the time interval between the second touch of iCub and the first one of the subsequent trial (about 44% vs 13% of trials, Wilcoxon Signed-Rank test Z=117, p=0.001).

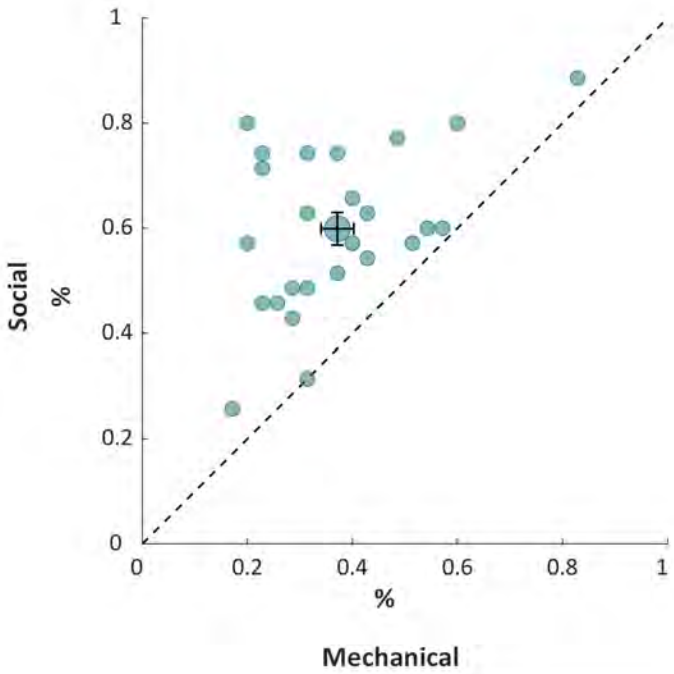

Figure 3.5 **Plot of the values of Godspeed subscale–Animacy.** Values are plotted for each participant in both mechanical and social conditions.

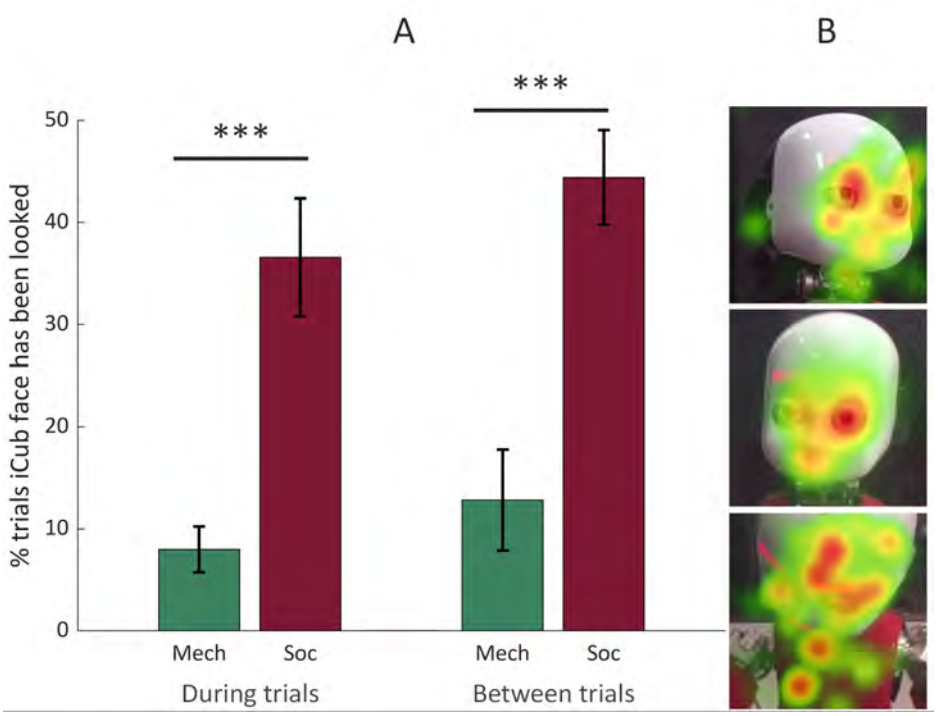

Figure 3.6 **Participants' gaze behavior towards iCub's face.** Figure A. Bar plot of the % of trials in which participants looked at iCub face during trials (tot trials = 66 trials) and between one trial and another (tot intervals = 65) in the mechanical and in the social condition. Figure B. Heatmaps of participants gaze on three representative snapshots referred to the mechanical condition (the one above) and to the social condition (the two below).

3.3.2 Context Dependency and perceptual errors

The main goal of this study was to understand the implication of a social scenario towards the use of priors in perception and to attempt a description of it using the Bayesian Model that, up to now, had been employed to describe perception only in individual scenarios [38, 195, 113]. Participants exhibited a significant degree of Context Dependency (regression to the mean) in the individual condition, with an average regression index of 0.446 (SD=0.133), significantly larger than 0 (one-sample t-test, $t(24)=16.8$, $p<0.001$, Cohen's $d=3.36$). Participants' perception was influenced by prior knowledge, leading to overestimating the shorter stimuli and underestimating the larger ones (see Figure 3.7).

In the two conditions with the robots, participants still showed a Context Dependency phenomenon (M=0.263, SD=0.175, one-sample t-test, $t(24)=7.86$, $p<0.001$, Cohen's $d=1.57$), although to a significantly lower degree, as proved in a paired sample t-test between the individual condition and the average values of the two robot conditions ($t(24)= 4.65$, $p<0.001$, Cohen's $d=0.93$). This general decrease can be partially due to the difference in the type

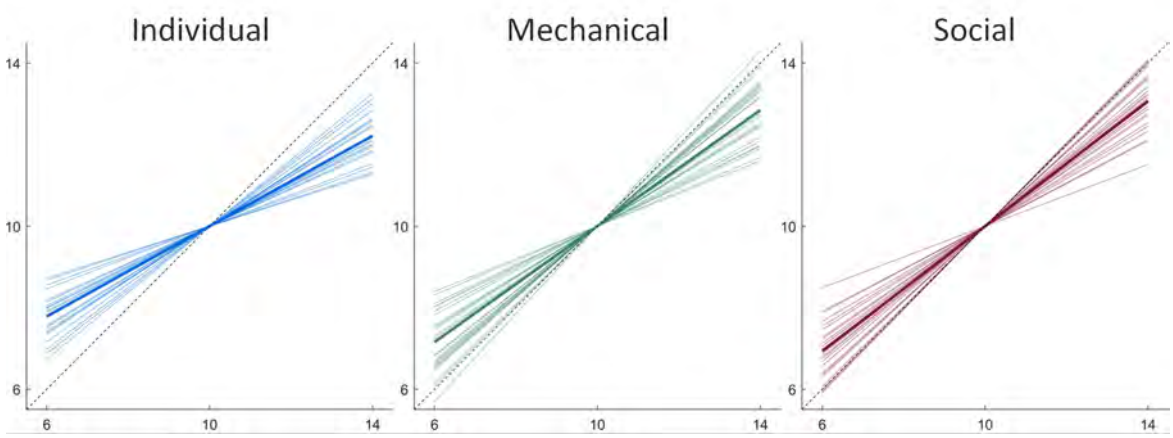

Figure 3.7 **Representation of the degree of Context Dependency in the three conditions**. Plots represent the slopes for each participant (thinner lines) and on average (thicker lines), resulting from the linear fit of the reproductions in the three conditions. The regression index is computed as the difference between the slope of the identity line (1) and the slope of the linear fit of data.

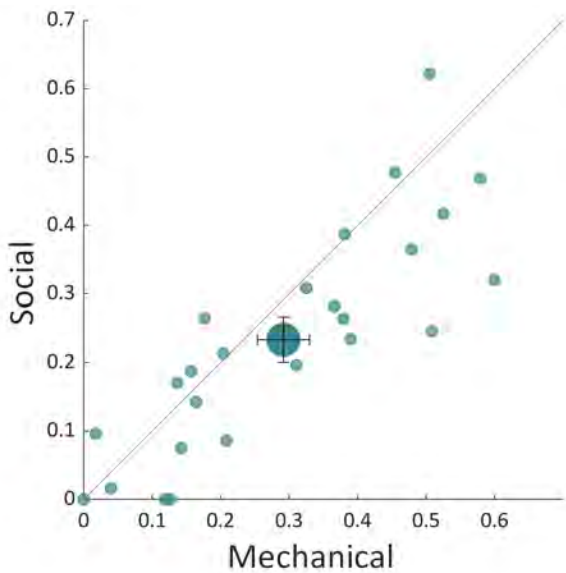

Figure 3.8 **Scatter plot of regression index values**. To compare the regression index in the two conditions with the robot, the smaller dots represent single participants in the mechanical and the social conditions, the largest one represents the mean with error bars calculated from the standard error of the two conditions.

of stimulation. In the individual condition, just two red disks represented the extremes of the length to be reproduced; whereas in sessions with the robot the whole arm motion was

visible, hence providing richer information. According to the Bayesian models described in [109], the presence of less sensory noise would yield a lower central tendency.

Considering the two sessions with the robot separately, participants exhibited a significantly lower degree of Context Dependency in the social-robot condition than in the mechanical one (mech: M=0.292, SD=0.183; soc M=0.234, SD=0.165), notwithstanding the sensory stimuli to be reproduced in the two conditions were identical (see Figure 3.7 and 3.8). A paired t-test comparing the two robot conditions revealed a significantly lower regression index in the social one (t(24)=2.92, p=0.007, Cohen's d=0.584). Therefore, results indicate that on average participants exhibited less central tendency when they were involved in an interactive context than when they were playing alone with a computer or with a mechanical device showing them the stimuli.

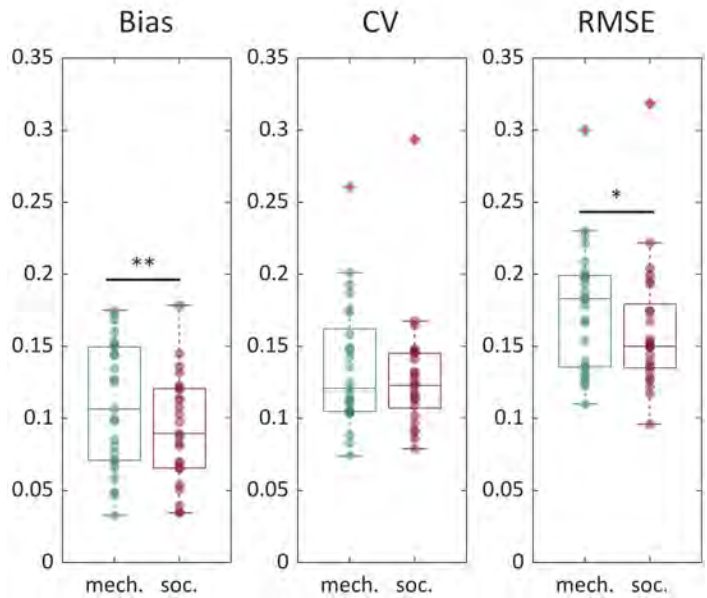

Figure 3.9 **Boxplot of perceptual errors.**. The values of perceptual errors (Bias, CV, and RMSE) in the two conditions with the robot (mechanical and social) are represented for each participant by circles. Perceptual errors have been normalized for the mean stimulus presented in the task (10 cm).

To verify whether previous experience interacting with iCub impacted the results, we divided all participants into two groups (if they had already performed experiments with iCub or not). A Mixed Model ANOVA with "condition" as within factor and "previous experience" as between factor did not reveal any significant effect of previous experience (F(1,23)=0.34, p=0.57) on the significant variation of regression index between conditions.

To deepen the understanding of the influence of social interaction on perception, we also analyzed the errors of reproductions, evaluating accuracy (bias), precision (CV), and total error (RMSE), as described in Section 3.2.6.1 (see Figure 3.9 and 3.11.A). We ran three Linear Mixed Effect Models, with the average error (bias, CV or RMSE) for each of the 11 stimuli as a dependent variable and the condition (Individual, Mechanical, Social) as a predictor. Furthermore, we applied random effects to the intercept at subject and stimulus levels. The random effect at the subject level has been applied to adjust for each subject's baseline level of error and model intra-subject correlation of repeated measurements. The random effect at the stimulus level served to model inter-stimulus variability in the error parameters. Random effects were submitted to the model in this order.

Firstly, we assessed the shift of both the sessions performed with the robot from the pure individual condition. We found a significant decrease of the bias both in the mechanical condition (Mechanical – Individual: B= -0.019, t=-3.74, p<0.001) and in the social one (Social – Individual: B= -0.033, t=-6.53, p<0.001). Such a difference could be partially attributed to the richer information of the stimulus in the conditions with the robot. With regards to the CV, it was not found any significant variation, neither with the mechanical robot (Mechanical – Individual: B= 0.0005, t=0.818, p<0.414), nor with the social one (Social – Individual: B= -0.00006, t=-0.099, p<0.921). Conversely, the RMSE was found significantly lower in the social condition (Social – Individual: B= -0.007, t=-3.525, p<0.001), but not in the mechanical one (Mechanical – Individual: B= -0.022, t=-1.112, p=0.266).

Since the two robot conditions were more comparable in terms of richness of information of the stimuli presented by the robot, we focus more specifically on the difference between them to assess the variation caused by sociality. Thus, we directly compared the errors of the two sessions performed with the robot with the three Linear Mixed Effect Models. Results revealed a significant effect of the bias (Social – Mechanical: B=-0.014, t=-2.784, p=0.005) and of the RMSE (Social – Mechanical: B=-0.015, t=-2.407, p=0.016), which resulted lower in the social condition. No significant effect has been found for the CV (Social – Mechanical: B= -0.005, t=-0.911, p=0.362) (see Figure 3.9).

To further understand the variation of perceptual errors we observed between the two tasks with the robot, we also performed a statistical analysis to find whether such a variation might be correlated with the variation participants revealed in how they perceived iCub's behavior in the two conditions. For this analysis, we used both the data gathered from questionnaires and the behavioral gaze data. Results of a Spearman correlation indicated that a significant negative association was verified between the variation in regression index (Δregression index: social-mechanical) and the variation in the value of anthropomorphism

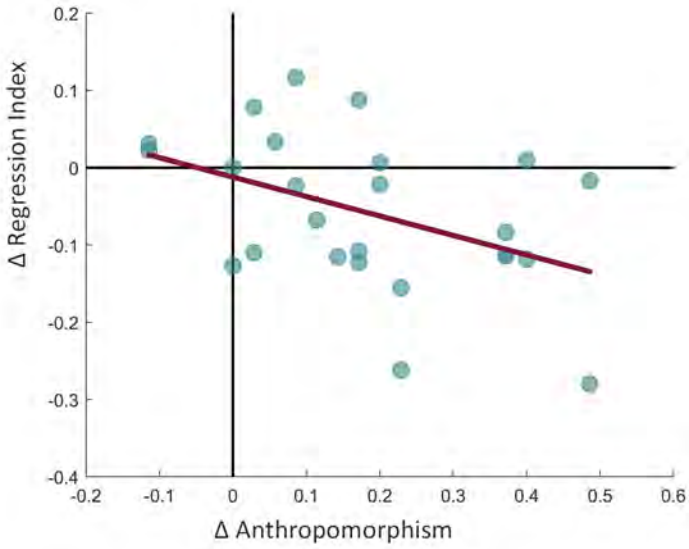

Figure 3.10 **Correlation Context Dependency - Anthropomorphism.** Individual variations of the regression index in the two robot conditions are plotted as a function of the variations in perceived anthropomorphism resulting from the Godspeed questionnaire.

(anthropomorphism: social-mechanical) ascribed to iCub in the two conditions: rs(25) = -0.446, p=0.025 (see Figure 3.10). The same correlation of anthropomorphism was also evident with the bias (Δbias: social-mechanical): rs(25) = -0.498, p=0.011. Such a result revealed that the robot aspect was the most critical feature of iCub that had an impact on perceptual data. The reason is that the Anthropomorphism scale includes questions about participants' impressions of the robot in terms of being fake - natural, machinelike - humanlike, unconscious - conscious, artificial - lifelike and moving rigidly – elegantly ([17]).

3.3.3 Simulation of the Bayesian Model

In Figure 3.11, data are plotted within the Bayesian framework that models Context Dependency as described in Section 3.2. The circles in Figure 3.11.A correspond to single participants' and average CV as a function of the corresponding Bias in the three conditions.

In terms of CV (precision), no difference is visible among the three conditions. On the contrary, considering accuracy, the bias of the three conditions decreases with this order: individual-mechanical-social condition. A similar pattern can be identified for the total error (RMSE), which can be seen as the distance from the axes-origin. The plot then clearly illustrates the results of the statistical analysis.

Starting from for 4 fixed values of σ_P (0.5 cm, 1.5 cm, 2.5 cm, and 3.5 cm) and from σ_L varying between 0 and 0.6, the continuous lines in the graph represent the model predictions

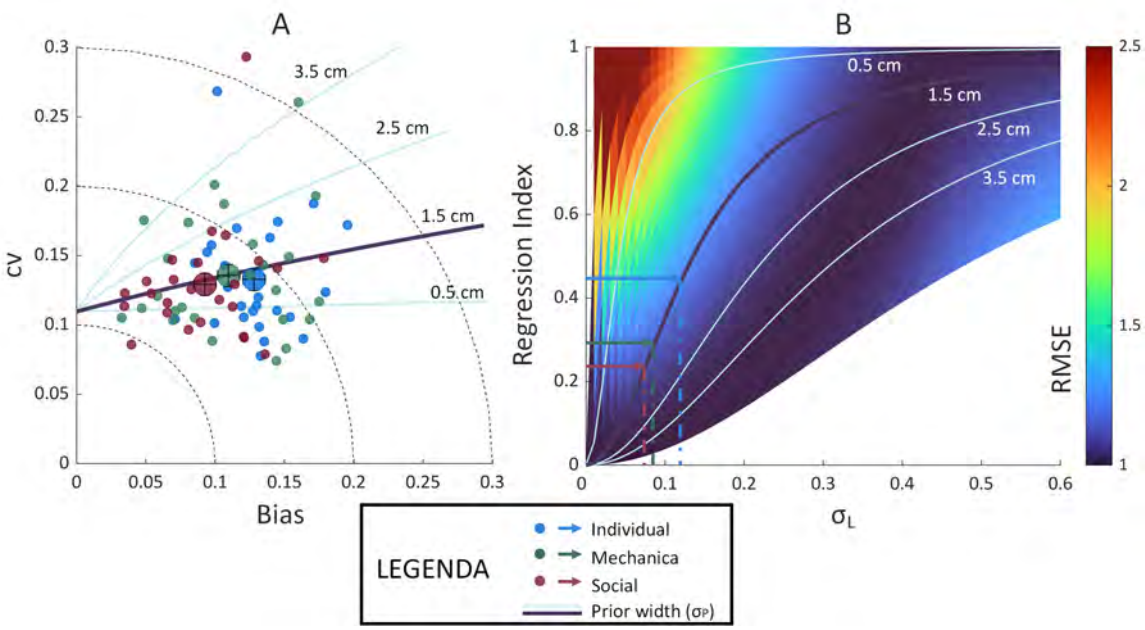

Figure 3.11 **Bayesian Model simulation.** Figure A shows the portioned perceptual errors in the three conditions: large circles represent the average normalized CV plotted against the average normalized bias with the error bars representing the standard error; small circles are individual participants. The four curves represent the prediction of the Bayesian model given a fixed value of σ_P (0.5 cm, 1.5 cm, 2.5 cm, 3.5 cm), which represents the weight given to the prior. Each curve has been plotted by varying σ_L (Weber Fraction) from 0 to 0.6. As in [38, 195], an additive fixed non-sensory motor noise of 0.12 has been added to CV. In Figure B, arrows represent the simulation of the model for σ_L, starting from the empirical data of the regression index and from the value of σ_P derived by the model (Figure A). In Figure B, it is also represented the value of RMSE simulated by the Bayesian model once given the regression index and σ_L and normalized for the minimal values of RMSE related to each value of σ_L.

for Bias and CV derived as described in Section 3.2.7 and normalized for the average stimulus (10 cm). As in [195], a further constant of 1.2 cm representing the non-sensory motor noise was also added to the CV. As shown in Figure 3.11.A, the results of all three conditions are predicted by the model with a σ_P of about 1.5 cm.

In Figure 3.11.B, the 4 sigmoid lines represent the model predictions about the relations between the regression index and σ_L for the same 4 fixed values of σ_P used in Figure 3.11.A. The background of Figure 3.11.B is color-coded to represent the different values of RMSE predicted by the model. The model predicts the highest values of RMSE when σ_L is low and the regression index is high: basically, in the case of an ideal subject who would have excellent eyesight but nonetheless relies heavily on its previous experience. Then, RMSE grows again when σ_L is high, but the observer does not regress enough to mitigate the error

caused by the weak eyesight. The lines derived from the model lie in the minimum of the RMSE as evidence of optimality.

We can assess where our data would lie on the model by considering the average regression indices measured in the three different conditions of the experiment as our ordinates. Assuming that the three conditions share the same prior width (1.5 cm, as derived in Figure 3.11.A), the model would predict that the perception in the three sessions was characterized by different σ_L.

As we mentioned above, σ_L is considered a function of the sensory threshold. Therefore, it depends on the observer's visual acuity or the richness of the stimuli's sensory information. A higher visual acuity – or more visible stimuli – corresponds to lower σ_L. A difference in the nature of the stimuli is indeed present between the individual conditions and both the robot ones. With the robot, the stimuli were provided by a gesture of the humanoid, whereas in the individual condition, they were indicated only by the red dots appearing on the screen. The richer sensory information associated with the robot action might therefore explain the lower σ_L in the robotic conditions. Conversely, between the Mechanical and the Social conditions, there was no difference in the sensory information since the robot's movement was the same in both sessions. If we were to impose an equal σ_L between the Mechanical and Social conditions – given that the participant's acuity does not change and neither the stimuli – the model would predict a lower prior weight (higher σ_P) for the social condition. However, this hypothesis would be incompatible with the measured CV and Bias in the social condition (see Figure 3.11.A).

In summary, the switch from mechanically-generated stimuli to stimuli generated by a social agent – though physically identical – led to a different degree of Context Dependency in our participants. However, a model, which predicts the level of integration of prior experience in perception by uniquely basing the estimation on the sensory acuity of the observer or, in turn, on the physical properties of the stimulus that can affect its visibility cannot explain, alone, the data collected.

3.4 Discussion

3.4.1 Context Dependency in social interactions

As humans, we adopt effective strategies to reliably perceive what is around us, to interpret others' behavior, and, as a sum of the two things, to interact and coordinate with them in a shared environment. To do that, we not only consider the information coming from our

senses but also build and use internal models coming from previous experiences that help us to cope with the uncertainty of information. If we just remain at a purely perceptual level, perception can be seen as an inferential process where previous experience influences the percept by acting as prior toward the incoming sensory information. But how do we use such priors when interacting with another agent? Which influence do they have on our perception, for example, on our levels of accuracy and precision? And what such an influence can reveal about the way humans perceive and share the environment with others? The idea underlying this study aimed precisely to start answering these questions.

The Bayesian model defined in Section 3.2.7 has been so far used to study prior inference mechanisms in individual contexts of perception. No parameter is present to assess the variation that a social scenario could bring to perception. Therefore, it should be verified whether descriptive models of individual perception can account for the change induced by sociality and verified in the variation of perceptual errors. To achieve this, it has been used a humanoid robot as a reproducible and controllable stimuli demonstrator. This solution could combine the rigorous protocol adopted in standard perceptual studies with an embodied interactive context. First, our results indicate that the perceptual phenomenon of Context Dependency occurs even in a social-interactive context, where a social robot shows stimuli. This means that humans employ their priors even in a social interactive context to perceive the world around them.

The hypothesis driving this work is that the brain puts in place mechanisms that might favor the emergence of Shared Perception, even at the expense of selecting a sub-optimal solution, if compared with an individual strategy. Our results show that we favor accurate estimation of a physical stimulus – if embedded in an interaction – rather than a stable, though less veridical perception, as the one normally derived by optimization in individual situations produced by the central tendency mechanism. In other words, in interactive scenarios, accuracy becomes more important than robustness to perceptual noise to allow for the successful completion of a cooperative effort. Therefore, the current perceived stimulation (e.g. the length of a movement) becomes less biased by the stimulus history (i.e., by the average of the lengths previously observed) and there is a minor effect of the central tendency strategy during the interaction.

The strategy of Context Dependency is put in place by our brain to cope with the uncertainty of sensory input and to reduce variability at the expense of accuracy [38, 195, 113]. Theoretically, the decrease in Context Dependency observed in the social robot condition could have been associated with increased response variability. But our results show that the interaction with the social robot kept a positive impact on perception. With the social

robot, participants demonstrated a significantly higher accuracy (lower bias) with respect to the interaction with the mechanical robot and to the individual condition, without having a negative influence on precision (CV). This implies that

- participants were more focused on each current stimulus they received from the social robot, as revealed by their higher accuracy
- participants were not distracted by its social behavior.

Even the overall error in reproduction measured by RMSE was significantly lower in the social condition.

Theoretically, in the Bayesian model, these results could be justified as a variation of two parameters: σ_L and σ_P. Nevertheless, as resulted from the simulation of the model with data, the shift between the social and the mechanical condition seems not to be explained either to an increased σ_L due to more visible stimuli or higher sensory acuity, either to an increase in σ_P that is a weaker prior. That being the case, the descriptive model based on individual perception does not account for the variation induced by the interaction with the social agent. Accordingly, in a more general model of Context Dependency, it is necessary to consider that the inferential processes of perception are a function of the social context, which could be described as reliance on one of the two sources of information: the current stimulus shared with the partner, or the private internal model built upon one's own experience. Thus, the perceptual mechanism of Context Dependency would also depend on the shared context of perception that may bring each partner to be more attentive towards the shared reality and to exploit less the private internal models about the world around.

This means that, given the exact same stimuli provided by the two robots, when interacting with the social robot, the inferential processes of perception are affected in favor of higher reliance on sensory information and a weaker dependence on the priors. Therefore, our results suggest the social interactivity of the context represents an additional factor modulating the integration of prior knowledge and incoming sensory stimulation.

3.4.2 Impact of robot's behavior

The effect of robot sociality on human distraction has been studied in different tasks and is an open question in the field of human-robot interaction. On this issue, in [116] it has been evidenced that the social behavior of the robot negatively affected the child's learning with respect to mechanical behavior. Authors hypothesized this could be due to distraction caused by the social robot or by a higher cognitive load induced by the social interaction.

Ingle et al. [100] found that, in a perceptual load search task, humanlike or anthropomorphic faces distracted participants in their task. In Spatola et al. [207], the authors showed that a threatening humanoid robot, but not a social one, increased the level of participants' attention during the Stroop task. From these studies, it seems therefore that the sociality of the robot might constitute a distracting factor in diverse domains.

With respect to this hypothesis, our results seem to go in a different direction. We found that in adults the social interaction with a humanoid robot, perceived by participants as more humanlike, likable, intelligent, and closer, did not affect human distraction, as suggested by the fact that the variability of responses (CV) does not increase, and the total error (RMSE) is even lower. Comparing the present study with related research on this issue, it is worth noticing at least three elements: the role of the robot, the cognitive load of the task, and the demographics of participants. Specifically, in the present experiment, the perceptual task was designed to be intrinsically interactive so that the robot was not only present in the scene as a distractor [100], or a tutor/instructor [207], but it rather had the role of stimuli demonstrator for participants. This could explain the reason why the robot did not constitute a distraction for participants. Moreover, our results might also be explained by the fact that the reproduction task of this study was not cognitively or perceptually high demanding. Lastly, the experiment was designed to be performed by a demographic of adult participants. Context Dependency has been already studied in visual perception of space in children [195], but only in the individual condition. Therefore, the question of whether a social robot distracts participants' perception is still open for this other age range.

From a comparison between the robot's behavior in the social condition and the mechanical one, the robot's gaze seems to play a significant role. The social session was indeed designed to establish mutual gaze with participants between one trial and another and precede the hand moving towards the point predetermined for the touch. The ultimate purpose was to strengthen the belief in intentional behavior in its human partners. On the contrary, when behaving mechanically, the robot directed its gaze in a static way toward a point that diverged both from the participant and the touchscreen. As it was viewed by Yonezawa et al. [256], the behavior of a robot responsive to their partners' gaze and establishing joint attention with them enhances both a favorable feeling of the users toward the robot and the users' belief of a favorable feeling of the robot toward them, if this behavior is also supported by the eye-contact reaction. Kompatsiari et al. [122] showed the positive impact of eye contact on human engagement in the interaction with the robot. Our results concur with these findings. The data gathered with the questionnaires highlight a substantial explicit preference for the social robot. They are also supported by the behavioral measures of participants' gaze. The

social robot's face was looked at more often both during and between trials revealing that the eye contact established by the robot after showing the stimulus was reciprocated and created a social context that was appreciated by participants.

As it has been explained in Section 3.2.4, we opted for explicitly priming participants about the robot's intentionality and social skills. Our aim was to assess the phenomenon of Context Dependency in a social context in comparison with a non-social situation. We then attempted to reduce the variability of participants' beliefs about the meaning of the robot's "mechanical" behavior. Given the humanoid child-like shape of the robot, in fact, we could not exclude that some participants automatically would have anthropomorphized the robot, also interpreting the mechanical behavior not as such, but rather as social and negative (unfriendly or apathetic). We tried to minimize this risk with a design that foresaw congruent explicit and implicit information about the social (or non-social) nature of the interaction. It is relevant to note that the combination of explicit priming with the implicit behavioral cues produced a significant difference in participants' impressions of the robot between the two sessions, but this difference was still very variable among participants.

Results also suggest that the change in perceived "anthropomorphism" of the iCub between the two conditions played an important role. Indeed, the more the robot was judged as having increased its anthropomorphism in the social condition, the less the perception in that condition was influenced by the statistical context, with respect to the mechanical one. So it seems that the more human-like the partner was perceived to be by participants, the less their perceptual strategy considered the previous stimuli, in favor of the current one. Since the robot was in both conditions a humanoid platform, moving its arm and torso according to biological motion rules, it was the combination of gazing, facial expressions, and speech, together with the explicit experimental framing, that drove this change in judgment, with no change in robot shape or its arm motion kinematics.

In general, it has been demonstrated that a robot can influence human attention [244], actions [235], and cognitive mechanisms [118], only by implicit behavioral cues. Considering these findings, we may hypothesize that the robot's behavior might alone impact perception as well, in particular modulating Context Dependency. However, the present work does not allow quantifying the relative impact of the robot's implicit cues and explicit priming. Now that the phenomenon has been proven, it will be interesting to verify in future studies whether either the robot's behavior or explicit priming alone could impact participants' perceptions.

3.5 Conclusion

The increased anthropomorphism and social intelligence attributed by participants and induced by all robot's behaviors seem thus to be the cause of a change in the perceptual schemes of the human interactant. In both conditions, the robot provides the stimulus to the human with the same biological movement of the arm. Still, only in the social session the perception ceases to be merely private for the human and becomes a perception of something shared with another agent: a Shared Perception. The context of Shared Perception influenced the entire perceptual process so the integration of priors with sensory information was modified in favor of a major influence of the latter. It seems like the human observers were prone to evaluate more what was currently happening. Therefore, in Shared Perception, what was weighted more was the shared source of information of the perceptual process, i.e. the current stimulus, rather than the private internal model, i.e. the prior. Perception becomes shared when another perceiving agent is considered, something that in our experiment could happen rather with the social robot than with the mechanical one. This seems to produce a change in our perceptual mechanisms in that others' presence or behavior affects the entire inferential process of perception from which our percept emerges.

Perception of something that is shared among two agents may therefore become shared itself. However, the multiple meanings related to the concept of "shared" requires clarification. What is shared among two agents is the real object to which perception refers and can be "shared" at least in two senses. In the first sense, two interactants can perceive the same (shared) stimulus coming from the environment. In this case, the real object of perception is shared because both agents perceive it simultaneously. Accordingly, perception becomes shared because the social context affects the way one agent perceives that thing. In the second sense, one can perceive what the other agent shows, i.e. disclosed (shared) by the other. In this case, the object of perception of one agent is what is shared by the other through an action, a bodily reaction, or an expression. Therefore, the observer's perception can be called Shared Perception because the observer, while perceiving something in the environment, incorporates into her/his perception the other's relation to that thing. In our study, both interactants were looking at the screen together. Also, it was the robot that provided the stimuli by showing the points on the touchscreen. It was, therefore, a Shared Perception in both senses. In the first sense of "shared", this means what the participants were aware of seeing together with the robot, whereas in the second sense, "shared" means rather what the robot was showing in each trial. It is true that also the mechanical robot showed stimuli to participants. However, its action was not unified to any apparent perception since it behaved

as a mechanical arm: it was not a perceiving agent. That is the reason why in this case it is not possible to talk about Shared Perception. Only the social robot, thus, established a Shared Perception, a particular relation between the two social agents which significantly affected the observers' perception, as the results of this study demonstrate.

The idea of a Shared Perception raises how critical the ability of self-other distinction might also be in perception. Self-other distinction refers to the ability to distinguish others' representations from ours and is a key mechanism in empathy and, more generally, in understanding others [129, 210]. In this sense, the awareness that one's perception differs from others makes humans adopt peculiar perceptual mechanisms associated with social interaction. The enhanced reliance on the current stimulus associated with the variation of Context Dependency and induced by Shared Perception might be one of these social mechanisms of perception. Moreover, given the connection between the use of prior knowledge and developmental disorders proposed by [167, 172], the study of how sociality affects prior integration into perception becomes even more compelling.

The paradigm of Shared Perception and the study of inferential mechanisms of prior integration may bring a double outgrowth. First, in human-robot interaction, Shared Perception becomes crucial because it is a means to explore to what extent humans are affected by robots and interact with them as social partners. Secondarily, it may promote the development of interactive machines designed to adapt to human abilities, and, therefore, enhance the outcome of the interaction. In several human-robot interaction scenarios, it would be indeed desirable to improve the quality of the interaction by reducing human perceptual errors caused by distraction, false prediction, and uncertainty of the sensory information. That is the case of collaborative robots in industries as well as robots in rehabilitation contexts. Both settings where gestures repetition, distraction, and uncertainty due to an occluded visual perspective or a deficit in sensory receptors, may adversely affect human perception. Nevertheless, interaction with social robots used to give information in public places or help older people in clinics and domestic environments, or even more with robots employed in developmental contexts, would be deeply enhanced if their design and behavior were conceived based on human social skills. So, to advance toward improved collaborations between humans and robots, it is still needed to deepen the human perceptual mechanisms and the way they work during interactions.

Chapter 4

A neural network analysis of Context Dependency during social interaction

4.1 Introduction

As described in Chapter 3, Context Dependency is a perceptual phenomenon revealing how human perception integrates sensory information and predictions about the external world. Prediction is a fundamental function of the human brain underlying various cognitive functions [158, 157], including visual perception [28]. Learning about the world by collecting experience helps us to process incoming visual stimuli in a more cost-effective manner, as we can reuse previous observations to make sense of new sensations. Predictive coding [66, 213] is a widely accepted neuro-cognitive theory that aims to explain human cognitive functions by prediction making. It claims that perception and sensorimotor responses stem from the brain's ability to constantly generate predictions about its environment and the internal states of the body. Substantial neuro-physiological evidence is consistent with the interpretation that prediction inference happens at all levels of perception [48]. Also, most actions could be explained as aimed at minimizing prediction error: from learning basic skills [80] to interacting with peers [26].

The main assumption of the predictive coding theory is that humans use near-optimal Bayesian inference, and draw their motor-sensory decisions from combining sensory information with prior experience. They then update their prior distribution with the new information and use the updated prior for the next prediction about the world. As described in Chapter 3, in Bayesian inference [109], the posterior perception depends not only on the values of the sensory inputs and priors but also on the precision of this information. Specifically, signals with low variance (i.e. high precision) affect the posterior more strongly whereas signals

with higher variance (i.e. a lower precision) are less taken into account (see Figure 3.4 in Section 3.2.7). This integration of prior and sensory information, depending on the precision of these two signals, improves the robustness to noise in the environment.

As we demonstrated, in previous research the extent to which human perception is influenced by priors, i.e. internal predictive models about the world, changes depending on the social context. As social agents, human perceptual processes are inherently shaped by social interactions (for a broader exposition, see Section 2.4). Nevertheless, the nature of behavioral changes induced by sociality and their underlying cognitive processes are still not well-understood [101, 91, 55, 45]. The question about the underlying neural mechanisms of a perceptual phenomenon such as Context Dependency cannot be easily answered using a behavioral experiment since it would require an analysis of the neural activation of the human brain – a challenging task given its complexity. However, one way to investigate the potential underlying mechanisms of the observed behaviors is by using a computational model that replicates the human behavioral data using a simplified neural system – an approach that is commonly used to investigate broad behavioral phenomenons which lack clear hypotheses applicable at a neural level [130]. Such neural network approaches, which replicate the human behavioral data using a simplified neural system, may provide a tool for exploring the role of various neural mechanisms on human perception and generating new hypotheses to be tested in neuro-biological as well as psychological studies.

From this perspective, we trained an artificial neural network on the human behavioral data from the user study reported in Chapter 3 to better understand the neural mechanisms underlying the condition-dependent variation of reliance on the prior that were found in the human behavior. Specifically, this study[1] focused on the mechanisms that play a role in how humans differentiate between individual and social task conditions. The neural network used for this purpose was originally introduced in [161] and integrates a recurrent neural network model that learns to make predictions about the world, functioning as an internal model, and a Bayesian inference module that combines sensory input and the predictions of the internal model based on the precision of these two signals.

Two experiments were conducted using this model. First, we manipulate the hyperparameters of the model to modify the network's reliance on sensory and prior information. This allowed us to investigate how such alterations affect the behavioral output of the network.

[1]The outcomes of this work have been published in [230] in collaboration with the Interactive Intelligence group at Delft University of Technology, the Artificial Intelligence department at Radboud University and the International Research Center for Neurointelligence (WPI-IRCN) at The University of Tokyo. In this study, I personally contributed to recording-managing data, interpreting-discussing results, and writing the final paper. I wish to thank Maria Tsfasman and Anja Philippsen who took care of the Methods and the Analysis, and whom I collaborated with for the discussion.

Secondarily, we analyzed the neural dynamics of the trained neural network to evaluate which mechanisms the network might be using to differentiate the three conditions using its neural encoding.

With this design, we aimed to answer the question of which neural mechanisms might explain the change of the reliance on the prior and sensory signals found in the social condition. Our hypothesis is inspired by the Bayesian perspective on predictive coding. Behavioral differences may be caused by an altered precision of the sensory and the prior signals. For example, a more precisely perceived stimulus would cause a sharper perception and, consequently, a higher reliance on the sensory input when performing perceptual inference.

4.2 Methods and Training Procedure

The computational model used is introduced in Section 4.2.1. Thereafter, the description of the training procedure is provided in Section 4.2.2 and the Network performance in Section 4.2.3.

4.2.1 The computational model

The computational model used in this study is made of two components. The first one is a stochastic continuous-time recurrent neural network (S-CTRNN) [155] that serves as the internal model which learns to make predictions about the world. The second is a Bayesian inference (BI) module that integrates sensory input with the priors generated by the internal model. This network model was first presented by Oliva et al. [161] and was used to predict how people and chimpanzees would perform a drawing completion task in Philippsen et al. [173]. We chosed this particular model since it both follows the principles of predictive coding and allows us to modify the precision of the model's prior as well as the precision of sensory perception.

The S-CTRNN network is able to recurrently predict the mean and the variance of the next time step of a time-dependent signal, where the mean is the estimated next value and the variance expresses the uncertainty of this estimation. As a higher variance means that the precision of the signal is lower, and vice versa, the estimated variance may also be described as inverse precision. Formally, given input $\mathbf{x}^t$, the S-CTRNN predicts the mean μ_{prior} and the variance σ^2_{prior} of the sensory perception of the next time step $\mathbf{x}^{t+1}$ (following standard

conventions, we denote scalars as x and vectors as $\mathbf{x}$. However, note that for this experiment, the input dimension D=1).

The context layer consists of 25 neurons which we found to be sufficient for well learning the 1-dimensional task. All network connections are linear mappings with weights and no bias terms. The input is mapped to the context layer via weights $\in \mathbb{R}^{1\times 25}$, recurrent weights are defined $\in \mathbb{R}^{25\times 25}$, and the context layer is mapped to the mean and to the variance output unit, respectively, via a weight matrix $\in \mathbb{R}^{25\times 1}$.

To train the network to reproduce human behavior, the backpropagation through time algorithm is used as described by Murata et al. [155]. Specifically, during training, which proceeds in epochs, the likelihood (expressed by the output mean and variance of the network) that resembles the human data is maximized by updating the network weights. In other words, the prediction error, scaled by the estimated variance, is minimized.

The likelihood L that is maximized and consists of two terms $L = \ln L_{out} + \ln L_{init}$. L_{out} is the likelihood that the network's estimated mean μ_{prior} and variance σ^2_{prior} account for the observed input $\vec{x}$:

$$\ln(L_{out}) = \sum_{t=1}^{T}\sum_{i=1}^{D}\left(-\ln(2\pi{\sigma^2_{prior}}^{t,i}) - \frac{(x^{t+1,i} - \mu^{t,i}_{prior})^2}{2{\sigma^2_{prior}}^{t,i}}\right), \tag{4.1}$$

where T is the total number of time steps (here, $T = 22$), and D is the dimensionality of the input vector (here, $D = 1$).

The term L_{init} is used as introduced in [155] and optimizes the distance between the activations of the recurrent layer, the so-called initial states.

$$\ln(L_{init}) = \sum_{s=0}^{S-1}\sum_{n=0}^{N-1}\left(-\ln(2\pi\sigma^2_{prior}) - \frac{(\mathrm{u}^{0,n}_{(s)} - \hat{\mathbf{u}}^n)^2}{2\mathrm{v}_{\mathrm{dist}}}\right), \tag{4.2}$$

where N is the number of neurons in the context layer (here, N = 25) and S is the number of initial states that the network should differentiate (here, one initial state per condition and per participant is used, resulting in $S = 3 \cdot 25 = 75$ initial states). $\mathrm{u}^{0,n}_{(s)}$ refers to the initial state (e.g., neuron activation vector at time step $t = 0$ for the s-th initial state of the n-th neuron). $\hat{\mathbf{u}}^n$ is the (learnable) mean of all initial states and $\mathrm{v}_{\mathrm{dist}}$ (set here to $\sigma^2_{init} = 1e7$) is the predefined variance of the initial states.

Initial states are required because the S-CTRNN is a deterministic system. Therefore, given one set of activations of neurons of the recurrent layer and a specific input signal, the network would, once it is trained, always produce the same output. We wanted to train

the model to replicate the behavior of the 25 participants in the 3 different experimental conditions in a single model, such that we could directly compare the way that different conditions and different participants were represented in the neural system. By representing different participants and different conditions with different initial states, the separation of different types of behaviors within the network dynamics can be achieved automatically during the training process. In this way, different participants and different conditions can be represented in the network with different neural dynamics, while reusing the same neurons and weight matrices. Specifically, the network is provided with the information of which training trajectory belongs to which initial state during training. Using the two likelihood terms, the network gradually differentiates the initial states during training. L_{init} defines a target variance σ^2_{init} that determines the desired variance between different initial states (see [155] for details). Generally, a higher variance between initial states leads to a stronger separation of the neural dynamics of different participants and conditions.

At each time-step, the output mean and variance predicted by the internal model is fed into the Bayesian inference module where it is combined with the raw sensory input and the corresponding precision (Figure 4.1) depending on the ratio of sensory and prior precision. Specifically, the mean and the variance of the posterior distribution is calculated as:

$$\sigma^2_{post} = \frac{(H_{sensor} \cdot \sigma^2_{sensor}) \cdot (H_{prior} \cdot \sigma^2_{prior})}{(H_{sensor} \cdot \sigma^2_{sensor}) + (H_{prior} \cdot \sigma^2_{prior})}, \tag{4.3}$$

$$\mu_{post} = \sigma^2_{post} \cdot \left(\frac{\mu_{prior}}{(H_{prior} \cdot \sigma^2_{prior})} + \frac{x}{(H_{sensor} \cdot \sigma^2_{sensor})} \right). \tag{4.4}$$

The distinguishing feature of this computational model is that it allowed us to manipulate the reliance on the prior and the sensory signal via parameters H_{prior} and H_{sensor} to simulate a stronger or weaker reliance on either the prior or the sensory input. These two parameters function as a factor that is multiplied with the variance of the prediction σ^2_{prior} or with the variance that is associated with the sensory signal σ^2_{sensor}.

During training of the network, $H_{prior} = H_{sensor} = 1$ is used such that the network learns to replicate human data. These parameters can later on be changed to higher or lower values to modify the reliance of the model on prior or sensory signals. Specifically, choosing $H_{prior} > 1$ increases the expected variance of the prior, leading the network to rely less on the prior. In contrast, choosing $H_{prior} < 1$ decreases the variance and causes the network to rely more on its learned prior while performing the task. H_{prior} and H_{sensor} can be set independently from each other to increase or decrease the precision of either prior or sensory

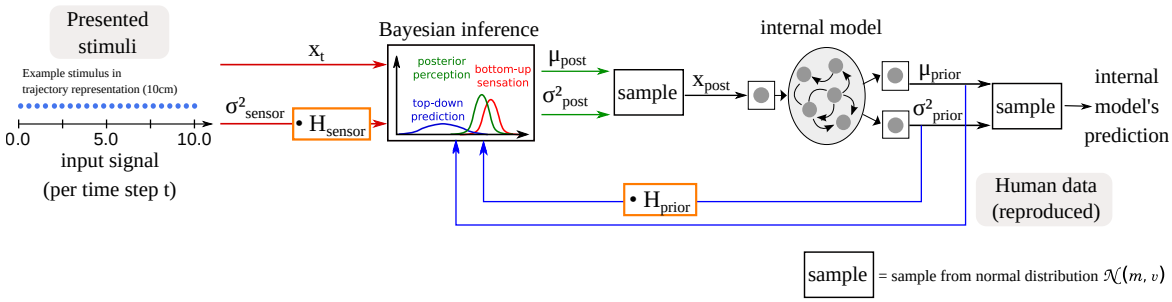

Figure 4.1 **An overview of the computational model used in the present paper**. A recurrent neural network serves as the internal model that learns to predict future time steps of a one-dimensional trajectory whose length represents the length of the stimuli.

information. Both parameters affect the ratio between the precision of sensory and prior information and, thus, have comparable effects on the model (an increase in prior precision has similar effects as the decrease in sensory information). Still, this study investigated such effects to understand how perception is affected, namely in how they determine the variance of the posterior.

4.2.2 Training the model to replicate human data

The main goal of this study was to verify whether the differences between individual and social perception between experimental conditions can be replicated by continuous modification of one parameter (e.g, prior reliance), and whether there might be multiple mechanisms causing behavioral differences. As a first step to investigate these issues (see Section 4.3 and 4.4), first the model had to be trained with the human behavioral data acquired in the behavioral study reported in Chapter 3. In this section, we describe how the network was trained and verify that the performance of the network replicates human performance with sufficient accuracy.

4.2.2.1 S-CTRNN training

In contrast to Oliva et at. [161] and Philippsen et al. [173], where the network was trained to directly reproduce the presented input trajectories (i.e. input equals output of the network), we trained the network by providing the stimuli presented to human participants as input while the output corresponds to the participants' reproduction of these stimuli. As such the training mimics human learning of the task as closely as possible.

The network was trained with all the data from the human experiment which involves the data of 25 participants who performed the task in three different conditions.

4.2.2.2 Training data

The training data were taken from the behavioral experiment described in Chapter 3. Since the S-CTRNN model is designed to learn the next time-step of trajectories, the lengths from the human data had to be modified into one-dimensional trajectories consisting of multiple time-steps. Each trajectory started at location 0 and ended at the particular length of the stimulus. Before using the data for network training, the trajectories were normalized such that all trajectory points fall within the range $[-1,1]$. Hence, after normalization, all trajectories start at -1. The representation of stimuli as a trajectory alters the setting from the human experiment where participants just pointed at the final position, but including intermediate points may also provide new opportunities. As we will see later in Section 4.4, this design allowed us to look into the length reproduction task as a dynamical process.

Both the presented stimuli and the lengths reproduced by participants were converted into multi-step trajectories in the same way. While the presented trajectories served directly as network input, the reproduced lengths were used for the prediction error computation during network training.

4.2.2.3 Training parameters

As motivated above, the model had different initial states for each participant and condition, resulting in $75 = 25 \cdot 3$ initial states. These initial states were automatically determined during training, using a high maximum initial state variance ($\sigma^2_{init} = 1e7$) to ensure that the neural dynamics of different conditions and participants were sufficiently separated from each other.

The parameters H_{prior} and H_{sensor} were set to 1 during network training while the number of neurons in the recurrent network layer was set to 25. The network was trained for 15000 epochs.

Ten networks were trained independently from each other, using different randomly chosen sets of initial weights. By investigating the performance of a set of networks, we can ensure that the results that we find are reliable and not caused by random effects.

4.2.2.4 Network behavior generation

Similar to the way that the human experiment was conducted, we tested the performance of the network by providing it with trajectories of different lengths. This test set corresponded to the data that was presented to the human participants. To generate the behavior for a specific participant and experimental condition, the corresponding initial state of the network was

used to initialize the activations of the recurrent network layer. Then, the network's output, given the input, was computed to generate the model's behavior. From the trajectories that the network produced in response to the presented stimuli, the reproduced lengths were computed as the absolute difference between the start and end points of the reproduced trajectory. A linear model was fit to the reproduced lengths in order to compute the regression index. Additionally, the neural activation history of the recurrent layer was recorded and used for neural representations analysis in Section 4.4. The resulting neural activation data consisted of the activation for each neuron of the model for each time-step, trajectory, participant, and condition.

4.2.3 Network performance

A comparison between the human behavioral data and the performance of a trained network for six randomly chosen participants in the three different conditions is presented in Figure 4.2. The x-axis shows the presented lengths, the y-axis the length reproduced by the human participants (left) or by the model when using the corresponding initial state (right). Lines in the right plot show the result of the linear regression that was performed in order to calculate the regression index.

It can be observed that the model is able to accurately replicate the mean of the human data. Note that for generating the results in this figure only the mean without the uncertainty was generated by the model to get a better impression of the model's behavior. Therefore, the variability of the human data is not replicated on the right side of Figure 4.2.

A direct comparison of the regression indices of the model behavior with the corresponding regression indices of the human behavior is shown in Figure 4.3 including the data of all ten networks. It can be observed that the model's behavior slightly diverges from human behavior, however, the large majority of stimulus replications accurately correspond to the regression index of the corresponding human participant. It can also be seen that in the individual condition, a stronger regression towards the prior is taking place than in the other conditions in the human data as well as in the model data.

Black dots in Figure 4.4 show the subject-wise difference between individual–mechanical, individual–social, and mechanical–social conditions, an important measure to visualize differences between conditions also used in Chapter 3. This distance is the highest for individual–social, indicating that the regression index is significantly higher in the individual condition compared to the social condition. The mechanical–social difference is smaller, but

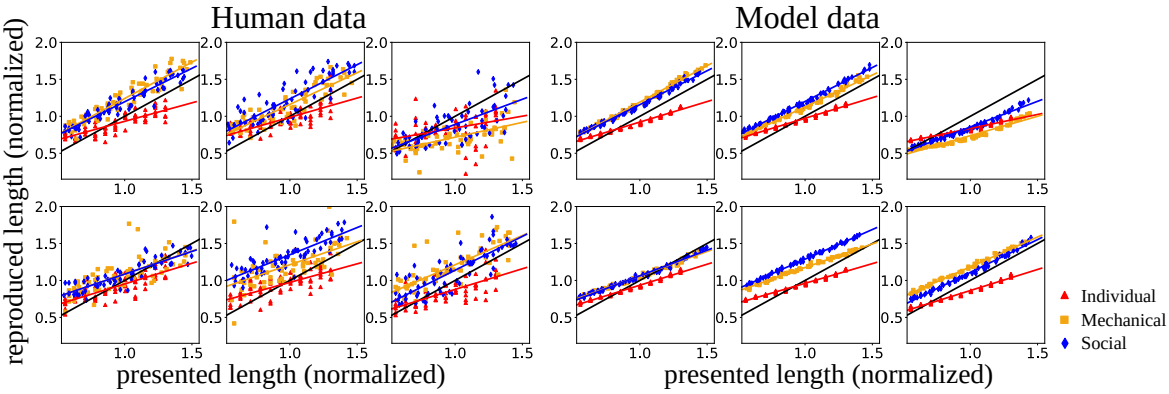

Figure 4.2 **Illustrative plots of reproduced lengths against presented lengths for original human data and model data**. Lengths were calculated in the normalized space of trajectories. Original human data (left) is compared with the corresponding mean predictions produced by one example network (right) for six randomly chosen participants. Lines in both plots correspond to the regression lines extracted from the human data or the model data, respectively. The black line shows the identity line.

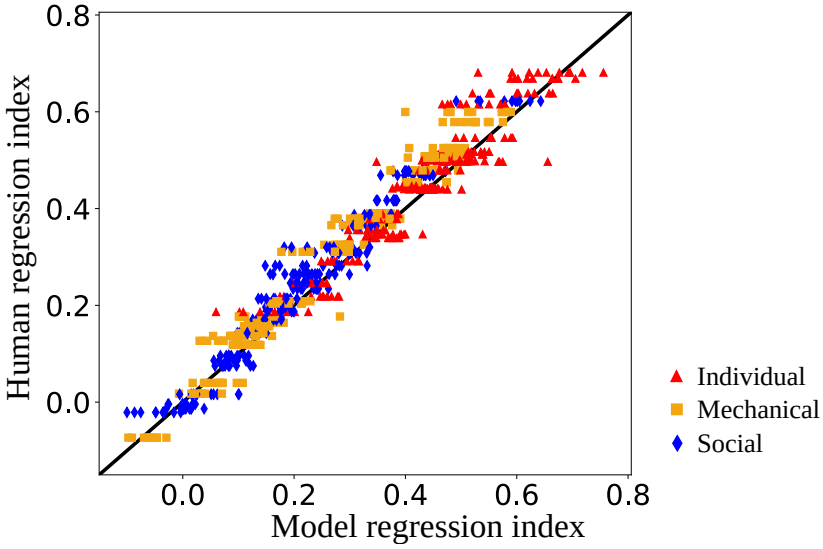

Figure 4.3 **Comparison between original data and model data in terms of regression index.** The regression indices of the human plotted against the regression indices of all trained networks for reproducing all training data.

significantly higher than zero, indicating that the regression indices of the mechanical and the social condition lie closer together.

Purple dots in Figure 4.4 show the same analysis conducted for the model results. It can be observed that the trends in the model behavior well replicate the human behavior, but the variability is slightly reduced in the model data compared to the human data. Specifically, the standard deviation of the model data is on average 7% smaller than in the human data.

Furthermore, there is a small significant difference between the model and the human data in the individual–mechanical condition difference.

The p-values, computed on all ten networks, are shown in Figure 4.4 in detail and were determined using linear mixed effect models, describing the subject-wise difference by either the conditions (e.g. individual–mechanical vs. individual social) or by the agent (i.e. human vs. model) with the subject ID as a random effect.

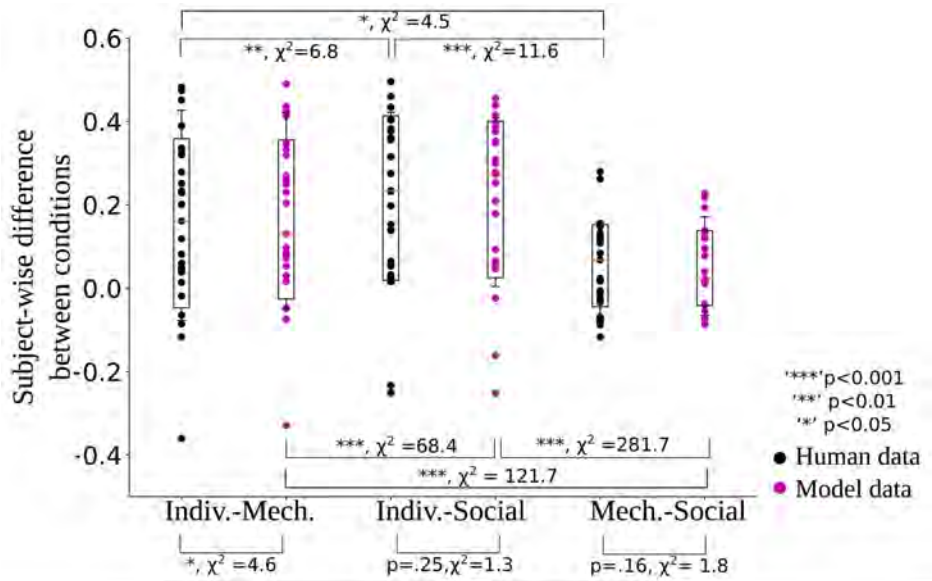

Figure 4.4 **Subject-wise differences between different conditions.** Differences are compared for human data (black) and model data (magenta) for one trained example network. Boxes indicate the mean, and 80% percentile of the data, fliers indicate the standard deviation. Model data reproduce the main trends of the data, but with slightly lower variability. The p-values were computed using the results of all ten networks, i.e. on 25 samples from the human participants, and 250 (= $10 \cdot 25$) samples from the models.

Overall, this analysis demonstrated that the model is able to replicate the important trends that are present in human data. Based on the trained models, we conducted two sets of analyses we call here Experiment 1 (section 4.3) and Experiment 2 (section 4.4). Experiment 1 aimed to answer the question whether it was possible to replicate the human results in different conditions with a continuous change of one parameter in the model. In short, Experiment 1 looked at how the model performed in the length reproduction task depending on its prior reliance. Experiment 2 investigated how the differences between conditions were represented in the neural activations of the network. It allowed us to look deeper into the mechanisms behind the differences in model performance and verify whether there are other processes at stake.

4.3 Experiment 1: Changes in the reliance on prior and sensory information

In the human experiment reported in Chapter 3, it was found that participants tended more strongly towards the prior in the individual condition, and more accurately replicated the stimuli in the social robot condition. The mechanical robot condition laid in between. This finding suggests that there might exist a continuum between the three conditions from the individual condition to the social condition via the mechanical condition.

The parameters H_{prior} and H_{sensor} of the computational model we are using here (see Section 4.2.1) can be used to implement such a continuous change as they modify the ratio to which sensory information and predictions are integrated while replicating the perceived lengths.

In this section, we tested the hypothesis that a continuous change of H_{prior} or H_{sensor} respectively can replicate changes in the human behavior between the individual, mechanical robot, and social robot condition. We first modified only H_{prior} (Section 4.3.1); then, we tested whether modifying H_{sensor} had analogous effects (Section 4.3.2).

4.3.1 Results Experiment 1A: Modifying the reliance on prior predictions

In the human experiment, the weakest reliance on the prior was found in the social robot condition. Therefore, our expectation was that when gradually increasing the model's reliance on the prior, a network behavior that was formerly replicating the social robot condition would produce behavioral results which would be closer first to the mechanical robot (with moderate increase of prior reliance) and then to individual conditions (strong increase of prior reliance). If this hypothesis was correct, it should have been possible to find values of H_{prior} such that the network behavior replicated the human behavior in the individual and mechanical robot condition, while only using the initial states of the social robot condition.

To test this idea, in this experiment, we used only a subset of the trained network dynamics, namely, the 25 initial states that are associated with the social robot condition. Then, we tested whether it was possible to replicate the results of the other two conditions by adjusting H_{prior}.

The network's behavior was tested by using a wide range of values between 0.5 and 0.05 for the H_{prior} parameter. For each of the different values of H_{prior} the network behavior was recorded. Similarly to Figure 4.4, subject-wise differences between conditions were

computed as a measure of how well the replicated lengths fit human data. Specifically, the difference was computed between the replicated length of the initial state of the social robot condition with $H_{prior} = 1$, and the replicated length of the initial state of the social robot condition with $H_{prior} = x$ where $x \in \{0.5, 0.45, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.09, 0.08, 0.07, 0.06, 0.05\}$. These values were selected in an iterative way based on how variable the behavior changed in a certain parameter region.

Figure 4.5a shows the median difference of all ten networks (different colors refer to different networks). The horizontal dashed lines in Figure 4.5 indicate the subject-wise difference between the social robot condition and the mechanical robot condition in the human data and the difference between the social robot condition and the individual condition in the human data. It can be observed that a stronger prior (i.e. a smaller value of H_{prior}) gradually increases this ratio, that is, with the increased prior reliance the produced lengths tend more strongly towards the mean of the data. A value of $H_{prior} = 0.4$ closely matches the social–mechanical difference of the human data, and $H_{prior} = 0.1$ closely replicates the social–individual difference of the human data.

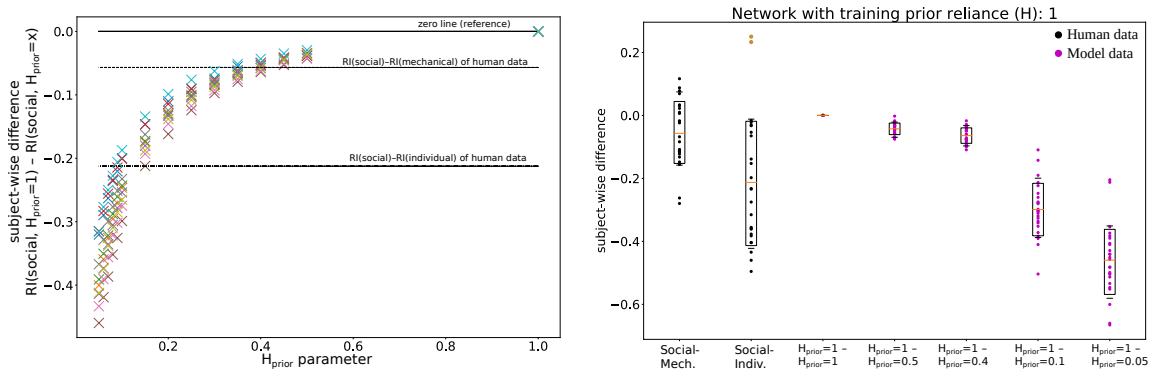

Figure 4.5 **Performance of the model as a function of H_{prior}.** Difference between the regression index of networks produced using the 25 initial states of the social condition with regular prior reliance ($H_{prior} = 1$) and the regression index produced with the same initial states using increased ($H_{prior} < 1$) prior reliance. (a) For all ten networks the median of the subject-wise difference is displayed. Horizontal lines mark the zero line, the average subject-wise difference in the regression index between the social and the mechanical condition in human data, and the average subject-wise difference in regression index between the social and the individual condition. (b) Detailed results including all subject data for a single network. The subject-wise differences between the behavior using social initial states of $H = 1$ vs. $H = x$ for different x values is displayed.

Figure 4.5b shows the subject-wise differences between conditions for a few selected values of H_{prior} for the data from a single network. This plot allows us to inspect not only the median but also the variability between different participants. It can be observed that

although the median for $H_{prior} = 0.4$ and $H_{prior} = 0.1$ match the median of the human data, the standard deviation is much larger in the human data. However, the further away the value of H_{prior} is from the standard value of $H_{prior} = 1$, the larger the standard deviation becomes. We tested statistically whether there is a difference between the subject-wise difference reproduced by the model in the different conditions and the corresponding human data. For this purpose, we used linear mixed effect models describing the subject-wise difference as a function of the identity of the agent (i.e. whether it is human data or model data) using the subject ID and the network ID as random effects. The subject-wise difference between $H_{prior} = 1$ and $H_{prior} = 0.4$ and between $H_{prior} = 1$ and $H_{prior} = 0.1$ showed no significant difference when compared to the social–mechanical difference or the social–individual difference in human data, respectively ($p > .05$).

The results demonstrate that it is possible to replicate the individual and the mechanical condition using the initial states of the social condition, i.e. we can switch from weak towards strong prior reliance. Theoretically, we could also go into the opposite direction, trying to modify the network behavior by moving from a strong towards a weak prior, i.e., replicate the mechanical and the social condition, starting from the tablet condition. However, executing the experiment showed that the subject-wise differences of the tablet condition did not change regardless of the H_{prior}. Specifically, even when changing H_{prior} to a value close to 0, the subject-wise difference remains the same. The reason for this finding is that the networks were trained to replicate human data and not to replicate the actual presented stimuli. Human subjects do not have perfect precision, thus, the human data that the network was trained with also does not reflect the actual presented stimuli. Therefore, the network is not able to achieve higher accuracy than the human subjects even if the attention is shifted to the sensory signal. Demonstrating the shift from a stronger towards a weaker prior, thus, is not possible with the current experimental design. In contrast, it is always possible to shift towards a more strong prior as this does not require any knowledge about the presented stimuli but is implicitly known in the model. Therefore, we focus in this section on demonstrating the shift from a weak to a strong prior.

4.3.2 Results Experiment 1B: Modifying the reliance on sensory information

Section 4.3.1 demonstrated that changing H_{prior} can replicate the behavioral differences between the conditions. This parameter can be intuitively interpreted as the inverse precision of the network's prior. However, modifying the inverse precision of the sensory input H_{sensor}

could yield similar results. To test whether a change in H_{prior} or H_{sensor} better explain the human data, we repeated Experiment 1A, modifying H_{sensor} instead of H_{prior}. As explained above, the result of the Bayesian inference is mainly affected by the ratio of H_{sensor} and H_{prior}, but the absolute values of the two parameters change the variance of the posterior.

To evaluate whether changes of H_{sensor} equally allow us to change the behavioral output of the network according to the human conditions, we selected values of H_{sensor} such that the ratio between sensory and prior precision is the same as in Experiment 1A. For example, setting $H_{prior} = 0.5$ leads to a ratio between H_{prior} and H_{sensor} of $0.5 : 1 = 0.5$. The same ratio of 0.5 can be achieved by keeping $H_{prior} = 1$ but increasing H_{sensor} to a value of 2. Thus, the corresponding value of H_{sensor} that produces the same ratio as the H_{prior} value that was used in Experiment 1A can be computed as $H_{sensor} = H_{prior}^{-1}$.

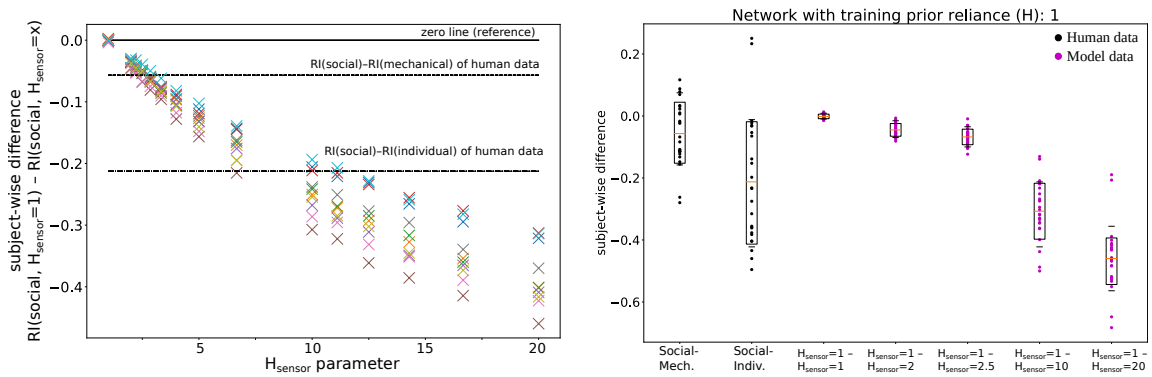

Figure 4.6 **Performance of the model as a function of H_{sensor}.** Difference between the regression index of networks produced using the 25 initial states of the social condition with regular reliance on sensory information ($H_{sensor} = 1$) and the regression index produced with the same initial states using decreased ($H_{prior} > 1$) sensory reliance. (a) For all ten networks the median of the subject-wise difference is displayed. Horizontal lines from top to bottom mark as indicated the zero line, the average subject-wise difference in the regression index between the social and the mechanical condition in human data, and the average subject-wise difference in regression index between the social and the individual condition. (b) Detailed results including all subject data for a single network. The subject-wise differences between the behavior using social initial states of $H = 1$ vs. $H = x$ for different x values is displayed for H_{sensor}.

The results are displayed in Figure 4.6a. Like in Experiment 1A, the figure shows the median difference of all ten networks (different colors refer to different networks) between the produced lengths observed with $H_{sensor} = 1$ and with H_{sensor} set to the values displayed on the x-axis of Figure 4.6. Again, the horizontal dashed lines indicate the difference between the social robot condition and the mechanical robot condition in the human data and the difference between the social robot condition and the individual condition in the human data.

While in Figure 4.5a the value of H was gradually decreased to increase the reliance on the prior signal, in Figure 4.6a the value of H_{sensor} is gradually increased to decrease the reliance on the sensory signal.

The results show a similar change of the difference with gradual modification of the parameter. The human data differences are replicated with $H_{sensor} = 2.5$ for social—mechanical and with $H_{sensor} = 10$ for social—individual. With these values, the exact same precision ratio between prior and sensory precision is achieved as with the corresponding values found in Experiment 1A. The corresponding plot of a single network 4.6b shows identical results to Figure 4.5b, indicating that in the present experiment a modification of H_{sensor} or H_{prior} lead to equivalent behavior changes.

We tested whether the difference between human and model data is significant for the individual parameter conditions analogously to the procedure described in Section 4.3.1. Also here, no significant differences were found for the above parameter values ($p > .05$), indicating that the model data well describe human data.

4.3.3 Discussion Experiment 1

The results of Experiment 1 indicated that reliance on the prior could account for the differences we saw in the behavioral differences between the three conditions. Specifically, we tested whether it was possible to gradually modify the network's behavioral output from weak prior reliance as it was found in the social robot condition of the human data towards a strong prior reliance as it was found in the individual condition of the human data. We found that a gradual shift of H_{prior} as well as of H_{sensor} could switch the network's behavior from the social condition to the other two conditions, indicating that all observed behaviors could be explainable based on the same underlying mechanism.

Notably, the same behavior could also be achieved by changing the reliance on sensory information instead of prior information. Further, while Experiment 1A and 1B could in principle yield differences in the variances of the behavioral output, no significant difference could be observed between the two mechanisms. Thus only the ratio, not the absolute values of H_{prior} and H_{sensor}, influenced the behavioral outcomes.

One reason why we did not find any differences depending on the absolute amplitudes of H_{prior} and H_{sensor} might be the fact that the task was too simple and thus easily learned by the network. A more complex encoding of the experimental data, which also takes into account the variability of the generated output could help to make differences between Experiment 1A and 1B visible. Here, the variance is estimated but not explicitly modeled in the data as

a sample is drawn from the estimated posterior distribution. Modifying the input encoding to explicitly model the variance of the signal, using for example population coding [77], could help to investigate whether differences between changes in prior and sensory reliance might exist. For the purpose of our investigation, however, the current implementation is sufficient as we were rather interested in the possibility to model the differences using a single parameter than in the differences between modifying prior or sensory precision.

In conclusion, Experiment 1 demonstrated that a gradual change of the reliance on prior or sensory information can replicate the changes that we observed in the human data. Therefore, it seems possible that human cognition makes use of the same underlying cognitive mechanism regardless of the situational context, but modifies this mechanism along a continuum to fit situational constraints. Specifically, the precision associated with the sensory and prior signal might be modified depending on the amount of social information that is present in the experienced situation.

These experiments demonstrated that changes of the precision of sensory and prior signals might be directly connected to the observed behavioral changes. However, this is only one possible explanation. In the following subsection, we explore the alternative hypothesis, namely, that there are fundamentally different cognitive mechanisms underlying the behavioral change observable between the three experimental conditions.

4.4 Experiment 2: Analysis of internal network dynamics

While the results obtained in Experiment 1 render it plausible that the same cognitive mechanisms might underlie the behaviors observed in all conditions of the human experiment, the differences among conditions in the human experiment might be caused by fundamentally different underlying cognitive mechanisms. For example, the difference between the individual condition and the two robot conditions seems to be of different nature than the change between the mechanical and social robot condition. In the first case, it was not a change in the social, but in the perceptual domain: whether the extremes of the presented lengths simply appeared on a screen or were indicated by the finger of the robot. The difference between the mechanical and the social robot condition, by contrast, was more subtle as the visual stimulus was identical whereas the change occurred in the social (or not-social) context of the task. Humans might thus use fundamentally different cognitive mechanisms to switch between the individual and the two robot conditions.

In this section, we investigated how the network model differentiated the three conditions, looking specifically at change in activations of neurons in the recurrent layer while replicating

the three conditions. Notably, differences between the experimental conditions were coded in our model only in terms of the behavior (i.e. the reproduced lengths). Differences in the way of presentation that were present in the human experiment (e.g. whether points appear on a screen or a robot touches the screen) were not explicitly modeled in the network. Thus, if we found that the network coded the responses in the three conditions differently, this indicated that this information should have been coded in the behavioral data of the human experiment, and the network automatically extracted them in order to solve the learning task.

Unlike Experiment 1, this analysis did not require us to modify any hyperparameter. Instead, we directly observed how the network self-organizes its structure to accommodate the dynamics caused by the three different experimental conditions. Since these are all trained within the same network, we could directly compare their corresponding network dynamics. The core question was thus whether the network dynamics reflected the differences between the individual and the robot conditions and between the two robot conditions, respectively, in different ways.

Therefore, we investigated how different conditions were represented in the internal activations of the neurons of the neural networks over the course of the trajectory (i.e. from time step 0 to time step 21). The activations at one point in time were a 25-dimensional vector containing the activation values of all the neurons in the recurrent network layer of the internal model. These vectors were generated for each time step, and for all human behavioral data, using the corresponding initial state of the participant and the condition in which the behavior was presented.

4.4.1 Results Experiment 2

An illustration of the network activations of time step 0 and time step 21 can be found in Figure 4.7. The activations are shown in the two-dimensional space generated via principal component analysis (PCA) from the original 25-dimensional vectors. In the left plot, the activations at time step 0 are shown, which correspond to the 75 initial states. Colored symbols label different experimental conditions, the black symbols and ellipses show the mean and the covariance of the three conditions. The right plot shows the activations at time step 21. Note that more points are visible in the right plot compared to the left plot because at $t = 0$ the trajectories still could not be differentiated depending on their length whereas this differentiation is reflected in the network activations at $t = 21$.

Qualitatively, it can be observed that the mean and the covariances are similar for the mechanical and social robot conditions, in the first as well as in the last time step. This result

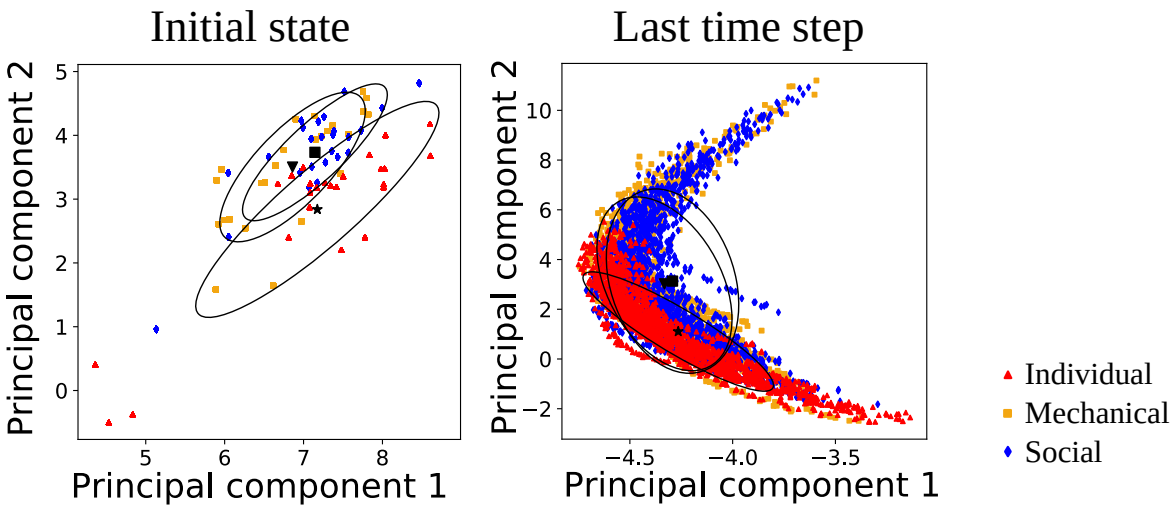

Figure 4.7 **Plots resulting from Principal Component Analysis.** The first two principal components of the network activation traces of one example network (capturing 83% of the variance), at the first time step (left) and at the last time step (right). The black symbols show the mean, ellipses the covariances of the points of the corresponding experimental conditions.

is to be expected because the behavior in these two conditions was more similar to each other. However, a difference between the first and the last time step can be observed in the covariances: in the first time step, the covariance is larger in the individual condition than in the robot conditions, whereas in the last time step the covariance appears relatively smaller in the individual conditions.

This covariance indicates how variable the internal activations are in each of the three experimental conditions. A higher variability at time step t indicates that the differences that arise between participants in this condition are coded more strongly in the network dynamics at this point in time.

Figure 4.7 shows only the results of a single trained network. To investigate whether there is a systematic change of variability over the course of time, we quantitatively measured the variability in the network activations of the three experimental conditions for all the ten networks across time.

To compute the variability between the activations of different participants in the network, we calculated the distances of the networks' activations within the three conditions as visualized in the scheme in Figure 4.8. In essence, activations were grouped into different categories depending on the length of the stimuli (eleven length categories were selected by identifying the most common presented lengths in the human data, namely, lengths which were presented more than 100 times during the whole experiment) and distances are computed only within the length categories. The reason for this procedure was that

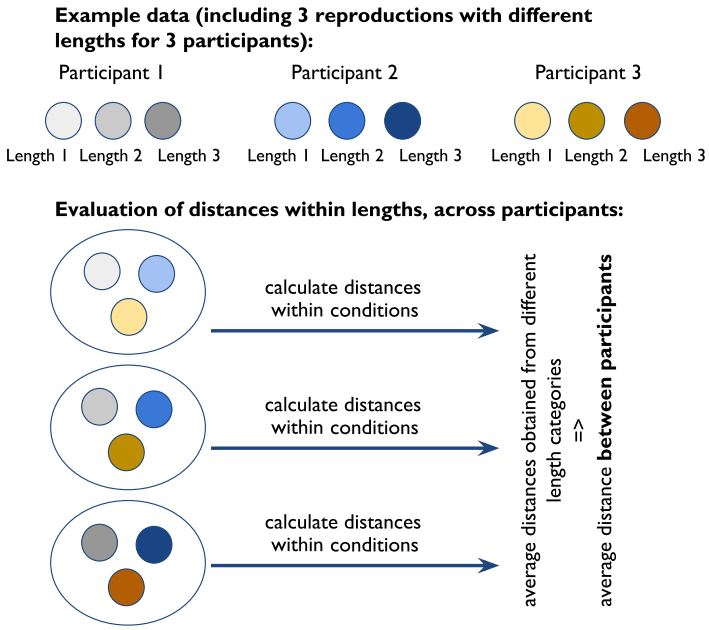

Figure 4.8 **Illustration of how the pairwise distances across participants were computed from the neural activation traces.** Each circle represents one trajectory of 25×22 where 25 is the number of neurons and 22 is the number of time steps. Data is split into 11 length categories and the pairwise distances within conditions are computed for each length category individually and later averaged, such that differences between lengths do not affect the final measure. The final measure, thus, shows for each time step the average distance between participants (see Figure 4.9).

we wanted to measure the differences in how different participants were represented in the network, but *not* differences in the reproduced lengths that also affected the network activations. Thus, the distances between all two activation vectors $\vec{x}$ and $\vec{y}$ of the same length category and experimental condition are computed as $1/N \cdot \sum_i(\sqrt{(x_i - y_i)^2})$. The results are shown in Figure 4.9. This plot shows the mean and standard error across the ten networks of the variability between activations of the same experimental condition. In line with the qualitative results in Figure 4.7, it can be observed that the individual condition has the highest variability in the beginning and the lowest variability in the end of trajectory generation.

The differences between the variability of the individual condition and the social condition are statistically significant in time step $t = 0$ ($p < .05$, R^{m2}=.16) as well as in time step $t = 21$ ($p<.05$, R^2=.16) when modeling the reproduced distance with linear mixed effect models and condition as fixed effect and network ID as random effect.

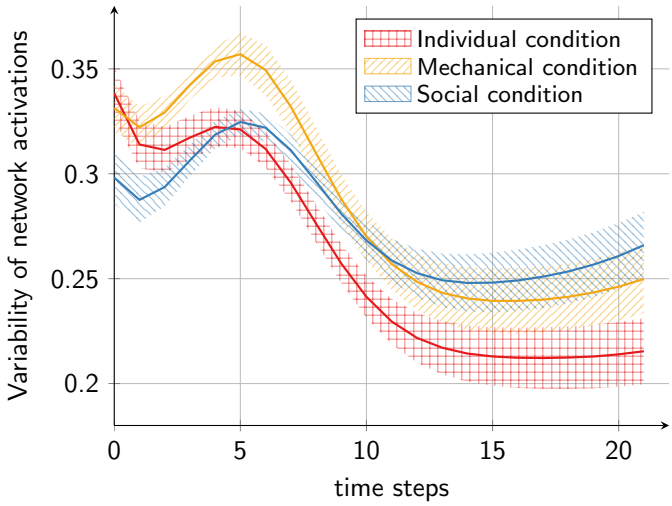

Figure 4.9 **Variability between the activations of different participants in the network.** Mean and standard error across networks of the average pairwise distances between the neural activation traces of the three different conditions (see Figure 4.8). Activations were normalized to [0,1] independently for each network beforehand.

4.4.2 Discussion Experiment 2

Results from this experiment provide access to the differences that exist among the individual, the mechanical and the social conditions in the generation of the trajectories. In particular, we suggest that the variability of network activations for the three conditions throughout the 22-time steps allows for a deeper understanding of how different the encoding of the network is across the three conditions for the entire generation process of the trajectories. Specifically, if the variability is high at time step 0, this indicates that the network mainly used differences in the encoding of the initial state of the network for differentiating the conditions when starting with the trajectory generation. On the contrary, if the variability is high at the last time step of the trajectory generation, this suggests that the differences between the conditions were mainly affected by the differences in the input data.

We observed that at the beginning, the individual condition shows higher variability than the social condition. Conversely, in the end, the variability is higher for the social as compared to the individual condition. For replicating the individual condition, the network mainly relied on information about the initial state, i.e., the network's prior information, whereas the social condition is affected more strongly by the input data that is presented during trajectory generation. This finding suggests that the neural network used different mechanisms to differentiate between the participants, depending on the condition.

While Experiment 1 demonstrated that the differences between the conditions could be explainable via a single unified mechanism, Experiment 2 suggested that the network might have used two fundamentally different strategies to encode the individual vs. the social condition, indicating that also multiple distinct mechanisms could be at play.

Firstly, the network relied on the differences in the initial states. In the experiment, this strategy could correspond to using context information about the perceptual task (top-down strategy). The second mechanism that the network used was the input signal (bottom-up strategy).

For what concerns the bottom-up strategy, in the behavioral experiment, the sensory information was richer in the case of the conditions with the robot compared to the individual condition (the robot's finger movements vs. a dot on the screen). It is, therefore, plausible that the participant relied more strongly on this richer information in the case of the two robot conditions. For what concerns the top-down strategy, the network shows a similar trend using more information about the input signal for generating the trajectories in the social condition compared to the other conditions (mechanical and individual conditions). This finding makes it plausible that such an interplay of two mechanisms could explain the behavioral differences.

Note that the input signal provided to the network input solely included the behavioral output, i.e., there was no difference in the richness of the signal in the computational study depending on the conditions. This limitation, however, is at the same time a strength of the computational model: the results hint at differences in the behavioral trajectories, although no confounding factors were present in the input signal. Still, it would be important in the future to verify this finding, extending the experiment to explicitly include factors such as the perceptual richness of the signal.

4.5 General Discussion

The aim of this study was to investigate how behavioral differences caused by differences in the social context could be replicated in a neural system, in order to generate hypotheses about the underlying cognitive mechanisms. For this purpose, we trained a neural network with human behavioral data of an experiment studying visual perception of space where three different conditions were tested ranging from an individual to a social task setting.

First, we demonstrated that the hyperparameters of the computational model that control the precision of the sensory and prior signal, respectively, can account for the differences among the experimental conditions (Experiment 1). Specifically, we found that altering the

precision of the prior as well as the precision of sensory input could replicate the behavioral differences between the three conditions: a stronger reliance on the prior, as well as a weaker reliance on sensory input, equally shifted the behavioral output of the network from the human behavior in the social condition towards the behavior in the individual condition, in line with the finding of Chapter 3 that participants tended more towards the mean in the individual condition. This finding makes it plausible that the same cognitive mechanism could be underlying the perceptual differences between the three conditions. Alternatively, different mechanisms could be intervening jointly in the same inferential process of perception.

The advantage of the network modeling study is that we could analyze the network's internal representation in order to understand how it performed the task at the level of neuron activities. Therefore, in a second experiment we analyzed how the differences between the conditions were coded in the neural dynamics of the network. In this second experiment, we did not artificially modify the network's mechanics (as in Experiment 1) but directly explored how the network internally represented and differentiated the three experimental conditions. The findings support the hypothesis of a plurality of phenomena affecting the visual perception of space. We found that the variations between the three conditions emerged at different moments in time, suggesting that different mechanisms are at play. At the beginning of trajectory reproduction, more information about the nonsocial conditions affected the network representation. At the end of the reproduction, the representation is strongly driven by the differences in the social conditions, potentially due to the richer visual input that was present in this task.

The findings of this second experiment indicate that the balance between sensory and prior information which we demonstrated in Experiment 1 only tells a part of the story. All three experimental conditions were differentiated in the neural encoding in ways that are intuitively explainable by the design of the human behavioral experiment (i.e. the richer sensory information provided in the robot conditions, see Section 4.4.2). The network solves the different conditions in different ways although it did not know what differentiated the individual and the robot conditions in the first place. This finding is interesting because it indicates that the human behavior alone was sufficient to let network dynamics emerge differently between the conditions, although the task design was exactly the same in all three conditions in the computational study.

4.6 Conclusion

As explained in Chapter 2, Shared Perception is an important aspect affecting our perception of the world. Here, as already emphasized in Chapter 3, the word “shared” could mean "disclosed to others" or “experienced in common” (see also [56]). Specifically, these two meanings referred to the stimulus that was communicated by the robot and experienced both by the robot and the participant. The framework of Chapter 2 described three elements involved in the process of Shared Perception: the environment (here the sensory stimulus), the self (the priors), and the other (the social context created by the robot’s social behaviors). The present work tried to understand how the social behavior of the robot could affect the relation between the other two elements producing a stronger reliance on the shared sensory information instead of private internal models.

Thanks to the use of a computational model where we can observe not only the behavior but also the internal dynamics of the network, in this study, we could add to the findings of Chapter 3 how the network came to the decisions it made. Specifically, the neural representation of the stimuli in the network allowed us to look into the time dynamics during the replication of the stimuli – something that remained hidden in the human behavioral experiment reported in Chapter 3. The proposed model simplifies cognition significantly but still might capture something important about Shared Perception, that is, how humans perceive their environment in a social context. The development of computational models for testing potential underlying mechanisms of specific behaviors found in human experiments, thus, may be an important means to form new hypotheses that may be tested in future experiments.

A further potential step for this research is to provide cognitive robotics with a computational model of Context Dependency in Shared Perception. This can be implemented on a robotic platform in order to endow it with a model developed on and "aware" of human perceptual mechanisms, able to take into account three different parameters: the sensory information, the prior, and the sociality of the context, which impacts on the other two parameters. Such socially perceiving robots might be used for further experiments in human-robot interaction to understand what social mechanisms would strengthen or reduce Shared Perception in similar perceptual phenomena.

Also, it could be interesting to repeat the experiment reported in Chapter 3 while looking at the dynamic changes of human behavior, by either changing the task design to a dynamic task or tracking human behavior over a longer time window.

Another important direction of future research is to strengthen the connection of computational study to the field of neuroscience. Such a stronger focus on computational studies for investigating neural phenomena is advancing significantly in recent decades, and a substantial amount of this work has focused on a topic that is also relevant for this study, namely, the relevance of top-down and bottom-up processing on human cognition [66, 213, 130]. Our analyses showed that human behavior in a social context might be affected by the precision of sensory and prior information and that two temporally separated mechanisms might be involved. Neuro-biological studies are required to understand which precise neural mechanisms are underlying such differences. There is in fact evidence of neuro-biological differences that can be measured in the human brain in a social context. The most prominent finding is that social context affects the concentration of neuromodulators in the human brain [45, 23]. Interestingly, neuromodulators also have been connected to the Bayesian framework. Specifically, studies suggest that neuromodulators might affect the reliance on prior and sensory information [43, 213]. Nevertheless, our study can only provide a potential explanation but not verify the neuro-physiological plausibility at this point. Further investigations are required to better understand the neuro-biology underlying social behavior in the context of a task like the one we investigate for gaining deeper insights into cognitive mechanisms of shared perception.

Chapter 5

A DL model for Addressee Estimation: a step towards an Addressee Estimation architecture based on Shared Perception.

5.1 Introduction

The etymological root of the word "communication" is the Greek "koinos", which means "common", "belonging to everyone", and "public". Communicating means, therefore, sharing something with someone, might it be a message, a thought, or an inner state. An act that inherently shapes the social world, binding together societies, social groups, and dyadic relationships. To properly be part of the social environment, each agent should understand some basic dynamics of communication, such as to whom a message is directed. A crucial element is, therefore, understanding who the addresser and who the addressee, or addressees, are. This ability is even more crucial for social robots, especially in those situations that go beyond the mere dyadic interaction.

The third research objective of this thesis (RO3) was developing a model for Addressee Estimation, which is the ability to understand whom someone is talking to, to foster robots' socio-perceptual skills based on the Shared Perception framework. This ability would bring tangible benefits for the robot in case of multi-party interaction, or in any scenarios where the robot is not considered the obvious focus of attention of other people. Grasping the others' addressee could make the robot able to understand implicitly expressed robot-directed commands, the level of engagement of all participants in a conversation, the social dynamics and roles in multi-party interactions, or the correct meaning of sentences that contains deictic

expressions (e.g., personal pronouns such as you, we, they, ...). Addressee Estimation is a concrete example of a capability that can be modeled on the Shared Perception framework and that reveals how integrating information coming from others (in this case, the speaker) can enable an augmented perception of the environment, such as discovering not-yet detected agents in the room.

To tackle this problem, this Chapter reported the development of a deep neural network for Addressee Estimation[1]. This model, which takes as input the speaker's non-verbal behavior and returns the addressee's position with respect to the robot yields the information coming from the other that the robot should integrate in order to achieve a complete Shared Perception. Therefore, it represents an essential step for a future architecture for Addressee Estimation that is outlined at the end of this Chapter and will be implemented in the iCub robot as future work.

For what concerns the structure of this Chapter, this first section aims to introduce the topic of Addressee Estimation and show its link to Shared Perception. After a more precise definition of the problem (see Section 5.1.1), previous works on the development of Addressee Estimation models will be presented (see Section 5.1.2). Eventually, the framework of Shared Perception, as defined in Chapter 2, will provide the structure for a possible implementation of an architecture for Addressee Estimation (see Section 5.1.3).

5.1.1 Definition of the problem

To achieve a definition of Addressee Estimation, the first step is to specify the meaning of the word "addressee". As Roman Jakobson stated, the addressee is one of the constitutive factors of verbal communication, together with the addresser, the message, the context, the contact, and the code.

"*The ADDRESSER sends a MESSAGE to the ADDRESSEE. To be operative, the message requires a CONTEXT referred to (the "referent" in another, somewhat ambiguous nomenclature), graspable by the addressee, and either verbal or capable of being verbalized; a CODE fully, or at least partially, common to the addresser and addressee (or in other words, to the encoder and decoder of the message); and, finally, a CONTACT, a physical channel and psychological connection between the addresser and the addressee, enabling both of them to enter and stay in communication.*" ([104], p. 21).

[1]This work started during a 3-month visiting period in the Cognitive Robotics Lab at the University of Manchester and was conducted in collaboration with Prof. Angelo Cangelosi and Dr. Marta Romeo. The outcome has been submitted at the International Joint Conference on Neural Networks (IJCNN) 2023 with the title "To Whom are You Talking? A DL Model to Endow Social Robots with Addressee Estimation Skills".

Hence, the addresser is the agent to whom the addressee intends to communicate the message. The addressee receives the message through a code and a channel connecting them to the addresser. In the relationship between the addresser and addressee, there is an element of intentional reference: the addresser refers to another agent as the addressee of its message and expresses this reference through a code that is understandable from the outside: through this code, the actual addressees can understand they were being addressed. Often, this reference is not only understandable by the addressee. There could be other agents, who albeit not directly addressed understand who the addressee is. In general, Goffman [81] divides the listeners of a speech into three categories:

1. the over-hearing listeners
2. the ratified participants that are not expressly addressed
3. the ratified participants that are expressly addressed.

This last group is that of addressees. For the communication to be successful, it is essential that also the other groups of listeners understand to whom the message is directed. As a consequence, this allows considering that within its communicative act, the addresser does not only communicate a message; it also communicates and expresses its addressee to all listeners. Behavioral studies demonstrated that the human expression of communicative aspects related to Addressee Estimation, turn yielding, and turn taking involves verbal, para-verbal, and non-verbal channels [206]. Specifically for Addressee Estimation it was proven that, beyond contextual and verbal information, the speaker's bodily cues, such as gaze and gestures, allow listeners to better understand the speaker's intentions [9, 103]. Our approach in developing our Addressee Estimation model was inspired by these findings. The scope of the present study was to implement a model for Addressee Estimation as an added social skill for robots, to enhance Human-Robot Interaction (HRI). To achieve this aim, we conceived Addressee Estimation as *the ability to understand an utterance's addressee by interpreting and exploiting non-verbal/bodily cues from the speaker.*

5.1.2 Previous works

Autonomous artificial systems that engage in multi-party interaction need Addressee Estimation skills because otherwise, it would not be possible for them to understand to whom a message is being directed. Addressee estimation models have been developed to go beyond the dyadic and robot-centric structure of human-robot interaction [110, 200] and enable robots to interact in more ecological scenarios [94]. Furthermore, other fields of research

for Addressee Estimation were the design of Smart Robotic Homes and Environments [180], human-behavior-aware conversational agents [143], meeting assistant agents [162], and other user-friendly systems, such as information kiosks [11] or artificial assistants [52]. Eventually, Addressee Estimation has also been used to enable systems detecting their partners' engagement in multi-party interaction [190]. This section intends to describe previous studies on the problem of Addressee Estimation and present methods and features used to approach its modeling.

The human expression of communicative aspects related to Addressee Estimation, turn-yielding, and turn-taking involves different channels. Together with contextual and verbal information, para-verbal (prosody and breathing) and non-verbal cues (gaze and gestures) allow listeners to understand the speaker's intentions [206]. Given these premises, previous studies have often dealt with the problem of Addressee Estimation by using a multi-modal approach. Jovanovic et al. [111] used an ad hoc retrieved dataset gathered on meeting groups of 4 humans to train a Bayesian Network and Naïve Bayes Classifiers with contextual, lexical, and gaze features and solve the task as classification of whom, among the four agents (or the entire group), was the addressee of each utterance.

Also, the AMI Corpus, containing data from 100 hours of meetings, served as a dataset to train models of several other studies [162, 65, 136, 137]. Frampton et al. [65] focused on the problem of the discrimination of the meaning of "you" as it was used in the utterances of the Corpus: (1) whether it was used in a generic vs. referential way, (2) with a singular vs. plural reference, and (3) by detecting its referent. To this aim, the authors trained a Bayesian Network classifier with visual (head location and orientation, speakers' focus of attention, and mutual gaze) and linguistic features (structural, lexical, and syntactic pattern of "you" utterance, relation with previous and next utterances, dialogue act features). Using the same dataset, Op den Akker et al. [162] treated Addressee Estimation as a binary problem from the perspective of each agent: "are you being addressed?" and trained different classifiers with data about the speakers' focus of attention, dialogue acts, and context of utterances (topic and role of speakers). To overcome the limitations of previous models, such as the specific position and the number of participants or the reduction of Addressee Estimation to a binary problem, Malik et al. [136] selected several features from the AMI Corpus (textual, contextual features, and focus of attention) to classify the role of the addressee in the meeting interaction as annotated in the corpus. In a later study, the same authors used similar features but added to the AMI dataset also the MULTISIMO Corpus, a multiparty multimodal dataset involving meetings of 3 participants [137]. In this way, authors trained different machine learning and deep learning algorithms to improve their previous results and

develop a real-time Addressee Estimation model, i.e., without information about previous addressees.

To avoid training their Addressee Estimation model on too specific situations and tasks, Le et al. [148] proposed a multi-modal deep learning model trained on images from the GazeFollow dataset [177] enriched with utterances generated by annotators, as if a person in the image uttered it. Considering images as snapshots from a robotic system camera, the authors trained their network with images, the cropped images of the speaker, their gaze direction, and the text of utterances to estimate if the addressee of the utterance was the person in the speaker's line of sight, the robot (the photographer), or others in the scene.

Research on Addressee Estimation, which directly involved artificial conversational systems and robots, mostly solved the problem as a binary classification. Bakx et al. [11] conducted multi-party experiments with two humans and an information kiosk. By recording participants with an external camera, they used a rule-based approach to classify whether the participant at the information kiosk was addressing the system or the human partner, given its focus of attention and the length of the utterance. Operating with the same scenario, Turnhout et al. [234] trained a Naïve Bayes Classifier to solve the same task. Katzenmaier et al. [114] designed a multi-party interaction with two humans (host and guest) and a simulated robot. They approached the task as a binary classification (host speaking either to the robot or to the guest) using visual data (automatically extracted head pose) and speech data (syntactic and semantic information and utterance length), both separately and combined. In a human-human-robot online interaction, Richter et al. [180] opted for a rule-based model taking as input the human's lip movement and the mutual gaze between the human and the robot (either combined or not) to understand if an utterance was addressed to the robot or not. After a dataset collection of multi-user human-virtual agent interaction conducted in Wizard-of-Oz, Huang et al. [95] trained an SVM classifier for a binary classification (robot addressed or human partner addressed) by giving as input several features related to prosody, utterance length, and head direction and equipped the virtual agent with a model for real-time Addressee Estimation. The work of Sheikhi et al. [200] relied on the role contextual information plays in Addressee Estimation. The authors used context about the utterance, the agents involved in the interaction, and the objects of interest in the environment to extract information about the speaker's and the human partner's visual focus of attention and to train a model to predict the addressee of each utterance in a binary classification task.

Table 5.1 **Previous Works on Addressee Estimation Models**

Paper	Training dataset	Task	Approach	Type of Features	Model Implementation
Jovanovic et al. [111]	Ad hoc retrieved Dataset (meeting groups of 4 humans)	multi-class classification (n subjects + group)	Bayesian Network Naïve Bayes Classifiers	linguistic, contextual, visual	Only software
Frampton et al. [65]	AMI Corpus	binary classification multi-class classification	Bayesian Network classifier	linguistic, visual	Only software
Op den Akker et al. [162]	AMI Corpus	binary classification "are you being addressed?	ML algorithms (best: Logistic Model Trees)	linguistic, contextual, visual	Only software
Malik et al. [136]	AMI Corpus	multi-class classification (n subjects + group)	Several ML algorithms	linguistic, contextual, visual	Only software
Malik et al. [137]	AMI Corpus + MULTISIMO Corpus	multi-class classification (n subjects + group)	Focus Encoding + ML + DL algorithms	linguistic, contextual,visual	Only software
Le et al. [148]	GazeFollow dataset enriched with utterances generated by annotators	multi-class classification (Line of sight entities, Photographer, Others)	CNN + LSTM	linguistic, visual	Only Software
Bakx et al. [11]	Data recorded in multi-party exp. (2 humans + 1 artificial agent)	binary classification (artificial agent or not)	Rule-based approach	linguistic, visual	Information Kiosk
Turnhout et al. [234]	Data recorded in multi-party exp. (2 humans + 1 artificial agent)	binary classification (artificial agent or not)	Naïve Bayes Classifier	linguistic, visual	Information Kiosk
Katzenmaier et al. [114]	Data recorded in multi-party exp. (2 humans + 1 artificial agent)	binary classification (artificial agent or not)	Bayes Classifier Multi-Layer Perceptron	linguistic, visual	Simulated Robot
Richter et al. [180]	Data recorded in robotic smart home (WoZ)	multi-class classification (robot or other elements of the smart home)	Rule-based approach	visual	Data recorded in Smart Home (model not implemented)
Huang et al. [95]	Data recorded in multi-party human-virtual agent interaction (WoZ)	binary classification (artificial agent or not)	SVM classifier	linguistic, visual, prosody	Virtual Agent
Sheikhi et al. [200]	Data recorded in multi-party HRI (2 humans + 1 robot): Vernissage Corpus	binary classification (artificial agent or not)	Logistic Regression	contextual, visual	Data recorded from the robot's sensors (model not implemented)

Addressee estimation has also been connected to other communication problems, such as turn-taking, and approached as a response obligation detection task. For instance, in a multi-party card game scenario, Johansson et al. [110] combined the turn-taking and addressee detection tasks into one decision: whether or not the Furhat robot has to take the turn. They used only automatically extracted features (voice activity, prosody, syntax, head pose, and card movements) and solved the task both as binary classification and as linear regression (gradual opportunity to take the turn). Horiguchi et al. [94] used an HRI dataset collected in ecological contexts (station and commercial building). The authors used a Long-Short Term Memory (LSTM) neural network combined with a Logistic Regression after extracting features related to vision (speaker's face), audio (Mel-frequency cepstral coefficients and Perceptual Linear Prediction), and text (word embeddings). In a multi-party HRI scenario, Romeo et al. [185] trained a convolutional neural network (CNN) model implemented on the Pepper robot over 4 days of interactions with humans for a three-class classification task. The robot was trained to predict the other agent's intention to interact by classifying visual images of the scenes according to three possible actions it could initiate in response to the situation. The robot could decide if starting the interaction, calling for attention, or waiting.

5.1.3 Shaping Addressee Estimation on Shared Perception

The problem of Addressee Estimation has been defined as the robot's task consisting of understanding to whom in the environment an agent (addresser) is addressing its speech. Starting from the framework of Shared Perception as it was structured in Chapter 2, this problem may be tackled through the triangulation of Shared Perception: the other, the self, and the environment.

In these terms, Addressee Estimation is the ability to interpret the addresser's intentionality and integrate it with internal models and perception of the external world. Therefore,

- The information from the other corresponds to the addresser's intentional relation to the environment.
- The information from the self embodies the set of the robot's internal representations of the environment and previous events. Among these, the robot's awareness of the external environment, its spatial memory of the disposition and number of agents in the environment, and prior/contextual information about the environment or previous events.

- The information from the environment, in this case, symbolizes the factual disposition and number of social agents in the environment that the robot can continuously check through visual perception.

To endow the robot iCub [147] with Addressee Estimation, we started designing a software architecture based on the Shared Perception framework. Endowed with audio-visual sensors, action generation abilities, modules for perceiving the environment, detecting persons, and retaining information about their location in the environment, the robot was missing the ability to interpret cues from the speakers' behavior about their addressee and to integrate the three above-mentioned sources of information. Two modules have been designed for this, but unfortunately, only the first one has already been developed.

1. The module "Addressee Position Estimation" (APE) is aimed at the interpretation of the non-verbal behavior of the speaker as a cue to localize the addressee. It has been developed as a deep-learning model classifying the addressee's position with respect to the robot starting from the addresser's face and body pose.

2. The module "Triangulated Addressee Identification" (TAI) is aimed at the final prediction of the addressee's identity and, although it has not yet been implemented, by design it will return the identity of the addressee by comparing:

    - the output of the APE module (information from *the other*)
    - the information that is available through the robot's spatial memory about the number and position of agents in the environment (information from *the self*)
    - additional information from perceptions of the environment allowed by action generation and further exploration (information from *the environment*).

The separation of the APE from the TAI module is intended to leave the first estimate of the addressee's position independent from the other sources of information (self and environment). The APE module interprets, therefore, the addresser's behavior and returns a first estimate of the addressee in terms of its position. In this way, the first reading of the speaker's intention may also be given in the lack of any hint given by prior experience or contextual and environmental information.

The TAI module is designed for a different task. This module takes the estimate of APE as input and integrates information coming from the robot's spatial awareness and novel perceptual knowledge coming from exploring the environment. At the crossroad of the three

sources of information (other, self, and environment), this module integrates, weights, and reads together different information, returning a final, triangulated prediction.

This Chapter explores the implementation of the APE module (see Section 5.2), which is the first crucial step for the entire architecture (see Section 5.4). The following paragraphs describe the methods (see Section 5.2.1), the results (see Section 5.2.2), and the discussion of this implementation (see Section 5.2.3). Subsequently, Section 5.3 conducts a preliminary assessment of the generalizability of the deep-learning model on the iCub robot. In Section 5.4, the design of the entire architecture and a discussion about its benefit for the robot in terms of Shared Perception is provided for future work.

5.2 Exp. 1. Development of “Addressee Position Estimation” (APE) model

In the APE module, we tackled the problem of Addressee Estimation as a classification task of the addressee’s position based on the non-verbal behavior of the addresser. To solve this task, we designed a deep neural network composed of two parts: a CNN network and an LSTM network. Three general principles guided the selection of the methods: Awareness of bodily non-verbal behavior, Temporality, and Suitability for ecological scenarios.

Awareness of bodily non-verbal behavior As Skatnze [206] explained, non-verbal information (e.g., head pose, gesture, contextual knowledge, prosody, proxemic, etc.) is crucial for Addressee Estimation. For this reason, given the embodied nature of Human-Robot interaction, the focus has been placed on the bodily aspects of non-verbal behavior. For instance, the gaze pattern, and hence the visual focus of attention, has been recognized as a crucial cue in previous work on Addressee Estimation [240]. More specifically, the head direction was often considered as a proxy in automatic models for Addressee Estimation in robots [200, 95, 110, 180, 206].

Following a similar direction, the model of Addressee Position Estimation implemented in this chapter is based on two non-verbal bodily information. The first one is the image of the speaker’s face. This image is used as information about the speaker’s head direction that, in turn, is a proxy for their visual focus of attention. The second is the body pose of the speaker, automatically extracted from the visual information coming from the robot’s cameras, represented by a vector of 18 key points and expressing the speaker’s whole body direction and gesture.

Temporality Another fundamental element that the design of our APE model considers is the temporal and ongoing nature of this kind of estimation. Addressee estimation is the ability to understand to whom an agent addresses a given utterance. Since an utterance can last from less than one second to several seconds, a reliable prediction cannot result from single, instant information. What is informative of the addresser’s intentional reference is the temporal sequence of its behavior. For this reason, the design and the training of the neural network of the APE model have been based on sequences of the addresser’s faces and body poses.

Another crucial point connected with temporality is the relation between duration of the utterance and estimate of the addressee. Only at the end of an utterance, all the elements to predict the addressee are available. However, it is also true that estimating the addressee before the utterance is completed is often possible and convenient. Accordingly, the classification task of APE module has been conceived so that predictions were independent of the length of the utterance and were regularly provided at set time intervals.

Suitability for ecological scenarios In implementing the APE module, the idea was to have a model as little constrained as possible by the interaction setting. With this objective in mind, the choice of data fell on those that could be automatically extracted through the robot’s sensors without the need for any external device. This principle guided the selection of the dataset to train the neural network. Therefore, a corpus (see Section 5.2.1.1) containing data recorded through the robot’s cameras was preferred to all those that exploited external sensors to record human movements and gaze. Thanks to this criterion, the model could be trained with data recorded from the ego-centric perspective of a robot, an aspect that will be extremely important when implementing the model on the robot to make it interact autonomously, without any help of external sensors.

The idea of overcoming the limits of dyadic HRI was another point inspired by the urge to leave the interaction setting as ecologically as possible. In real-world scenarios, the dyad addresser-addressee is often broken by other elements taking part in the interaction. The third element(s) can be represented by other agents as potential partners of the interaction or even by objects in the environment the two partners are talking about. In developmental psychology, interactions in which the infant and the adult interact with or through an object are called "triadic". Accordingly, a triadic interaction between the speaker and the addressee occurs when the addresser talks to the addressee about something in the environment, might it be an object or another agent. This situation, which frequently occurs, modifies the gaze

pattern and the body posture of both interactants [206]. Therefore, a second preference criterion for selecting the dataset was that it also included triadic interactions with objects.

5.2.1 Methods

5.2.1.1 The dataset

Figure 5.1 **Illustrative frames from Vernissage Dataset.** Examples of multi-party HRI data recorded from the Nao robot’s cameras.

Rather than creating a custom dataset from scratch, we evaluated if those already existing and publicly available answered to the requirements previously identified. Jayagopi et al. [107, 108] created a synchronized multimodal corpus of multi-party interactions in which two humans conversed with a Nao robot (Aldebaran robotics). During the interaction, the robot asked participants to present themselves to the group, showed them – in the role of an art guide – some paintings on the walls of the room, and asked them some questions about the paintings. Moreover, participants sometimes had to discuss together before giving the answer or could comment with their peer about what was going on. All recordings took place in the same room, where participants were not required to keep a specific absolute position, although most of the time, because of the configuration of the room and the interaction, they were standing in from of Nao and had a relative position to each other: one on the left, the other on the right side (see Figure 5.1 for some examples). Dialogues about the paintings on the walls ensured the presence of triadic interactions with participants that, when describing

the paintings, focused their attention on them or directly got closer to see them better. In this way, although the dialogue was controlled by the questions of the robot, the scenario granted sufficient flexibility and spontaneity to human behavior.

Yet another positive feature of the Vernissage Corpus was that the authors recorded the interactions from the cameras of the Nao robot, positioned in its eyes. In this way, the implementation of the APE model could be based directly on data provided by a robot from an ego-centric perspective and hence more easily ported onto other robotic platforms. As a side effect, data resulted noisier because of the movement of the robot while, for instance, nodding or turning its head, but this could also be an advantage in training a model suitable for ecological scenarios. For what concerns sensors, the robotic platform used for recording the Corpus had a modified head containing improved cameras located in the eyes of the robot (for further specifications, see [107]). Only a single camera was used for recording, with a resolution of 640x480 pixels, at a frequency of about 15 fps (mean) and YUV422 color mode.

The Vernissage Corpus is manually annotated to have a Ground Truth also for information about addressees and utterances. More specifically, the addressee has been annotated as “the person or group of people to whom a speech utterance is intended to” [107]. Annotation has been made by one coder, whose reliability has been positively tested through annotations from another coder on 30% of the annotations. Information about the addressee was annotated each 40 msec, starting from merged video and audio. Five different labels were used: "ROBOT", "RIGHT", "LEFT", "GROUP", and "NOLABEL". "RIGHT" and "LEFT" refer to the person at the right and the left of the robot, "GROUP", means that both the robot and the other agent are addressed, whereas "NOLABEL" indicates a time interval of silence or laugh. For annotation of utterances, silence and speech segments have been automatically detected and then manually controlled and adjusted.

5.2.1.2 Features selection and pre-processing data pipeline

The data chunk on which the neural network for Addressee Estimation has been trained is a sequence of 10 frames of face images and vectors representing body poses. The process to obtain the chunks from the Vernissage Corpus involved the following five steps:

- Division in utterances
- Extraction of body poses and face images
- Aggregation in sequences

- Data augmentation
- Body pose shifting.

Division in utterances The dataset comprised recordings of 10 interactions between two humans (different for each interaction) and the Nao robot. Video clips were trimmed according to the speech detection annotations and extracted from the recordings of the 10 interactions, leaving out the frames labeled "silence’. The rationale was to eventually keep only chunks in which participants were actually speaking. Since no verbal and semantic information has been used for the task, utterances have been defined only based on speech detection. Therefore, utterances were considered as the time intervals in which an agent continuously speaks without silence being detected, that is, utterances that are not interrupted by silence pauses longer than 0.08 sec.

Extraction of body poses and face images Firstly, utterances were divided into frames of 0.08 sec. Then, using OpenPose [33], vectors of 2D coordinates for body poses were extracted for each frame for all participants in the image. For the extraction, the OpenPose COCO body format was adopted, which predicted the x and y coordinates for 18 key points of each person (5 for the head, 3 for each limb, and 1 for the torso). Coordinates ranged from -1 to 1 for both axes. Given the poses, from the five coordinates of each person’s head-key points, a square-size cropped image was obtained by the original frame, resized at 50x50px. Body poses and face images were labeled as "speaker" or "other’. During the interaction, both human participants played the role of "speaker" and "other’. Since the Vernissage Corpus comprised 10 interactions, each with two possible speakers, 20 instances of speakers were available.

Aggregation in sequences The following step consisted of aggregating frames in sequences of 10 consecutive body poses and face images. Sequences, thereby, resulted being 0.8-sec portions of the utterances. From this aggregation, the data chunk of the dataset consisted of sequences of 10 body poses and 10 face images. Since the objective was to train a network with only data from the speaker, the speaker’s sequences were saved separately from those belonging to the other participant. All sequences were annotated with the addressee ("ROBOT", "LEFT", "RIGHT", "GROUP") and received an ID connected to the interaction, the speaker, the original utterance from which they were extracted, and the chronological order in which they appear within the utterance. Figure 5.2 shows an illustrative sequence, whereas Figure 5.3 shows the difference between sequences and utterances.

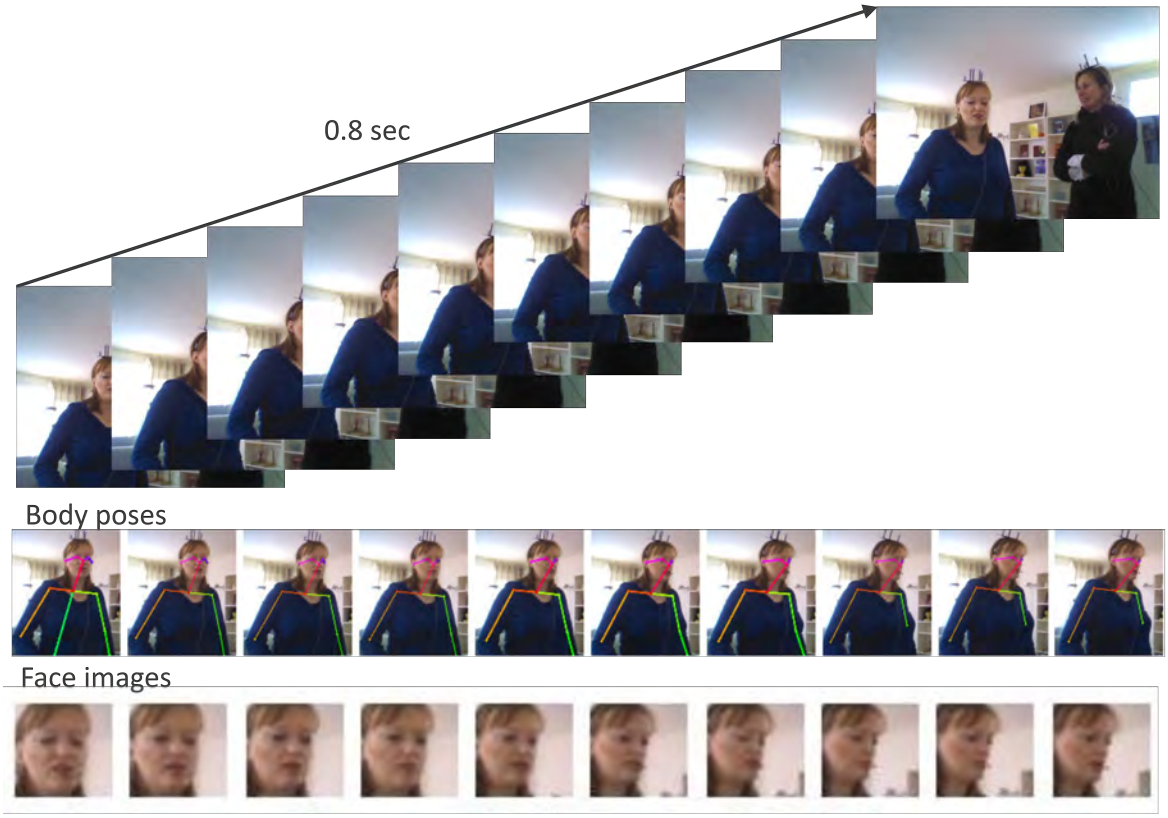

Figure 5.2 **Illustration of a sequence.** Aggregation of frames in a sequence of 0.8 sec. and extraction of body poses and face images.

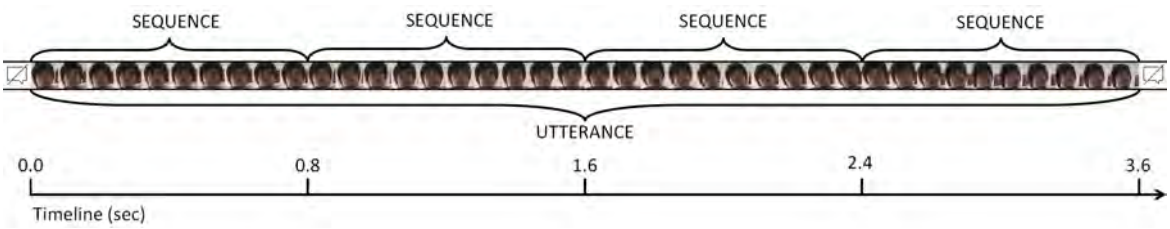

Figure 5.3 **Illustration of an utterance.** The utterance is partitioned into sequences of 0.8 sec. Utterances were defined as speech intervals addressed to the same addressee and delimited by silence. Each utterance comprised at least one sequence.

Data augmentation Given the purpose of the study (classifying the addressee’s position), we selected three labels: "LEFT", "ROBOT", and "RIGHT", leaving out the label "GROUP" that had no precise reference to the position of the addressee. The interaction scenario, with the Nao robot asking questions and managing the interaction, provoked an imbalanced representation of classes in the dataset, with a prevalence of sequences labeled as "ROBOT" (addressed to the robot). With the double objective of augmenting the dataset and balancing

the number of sequences, all frames labeled as "LEFT" and "RIGHT" (and accordingly, the body poses and the face images extracted by them) were flipped, and their label inverted (from "LEFT" to "RIGHT" and vice versa). As a consequence, "LEFT" and "RIGHT" data have been doubled, and the resulting dataset was composed of 18190 speaker’s face images and body poses partitioned in 1819 sequences: 529 for "ROBOT", 645 for "LEFT", and 645 for "RIGHT".

Body pose shifting A final step involved data transformation for all the body poses to prevent bias during the training process. As it appears from Figure 5.1, participants at the left of the robot never spoke toward a left addressee, and participants at the right never did it toward a right one. Therefore, even though participants could mildly move, the coordinates of their bodies could bias the prediction of their addressee. To overcome this issue, for each sequence, the 10 body poses were shifted along the x-axis of a random measure ranging from the two extremes of the image.

5.2.1.3 Architecture Design

The core objective of this study was to develop a model to estimate the position of the addressee. The task has been approached as a three-class classification of the addressee position starting from the addresser’s embodied non-verbal behavior. The position of the addressee was considered from the ego-centric perspective of the robot, taking the addresser as a reference. Therefore, the addressee’s position could be classified as "LEFT", in case the addressee was at the left of the addresser (from the robot’s perspective), "RIGHT", in case the addressee was at the addresser’s right, or "ROBOT", in case the addressee was the robot.

This configuration allows the model to extract features through convolution and then support learning temporal sequential patterns with LSTM cells. The deep learning model developed for the task is a CNN + LSTM hybrid architecture. Previous works combined CNNs with an LSTM final layer. For instance, Subramaniam et al. [217] used this combination to train a model for classifying first impressions of personality. Romeo et al. [186] used a similar architecture to predict apparent personality from body language cues for human-robot interaction. Moreover, Ullah et al. [233] integrated convolutional and LSTM layers for action recognition from videos, while Nakisa et al. [159] developed a multi-modal neural network with convolutional and LSTM layers for emotion recognition through physiological signals.

The model was also designed to exploit and integrate both visual modalities: the face images and the body pose vectors. Face images and body pose vectors pass independently in two parallel streams of convolutional layers. Consequently, the two embeddings received as

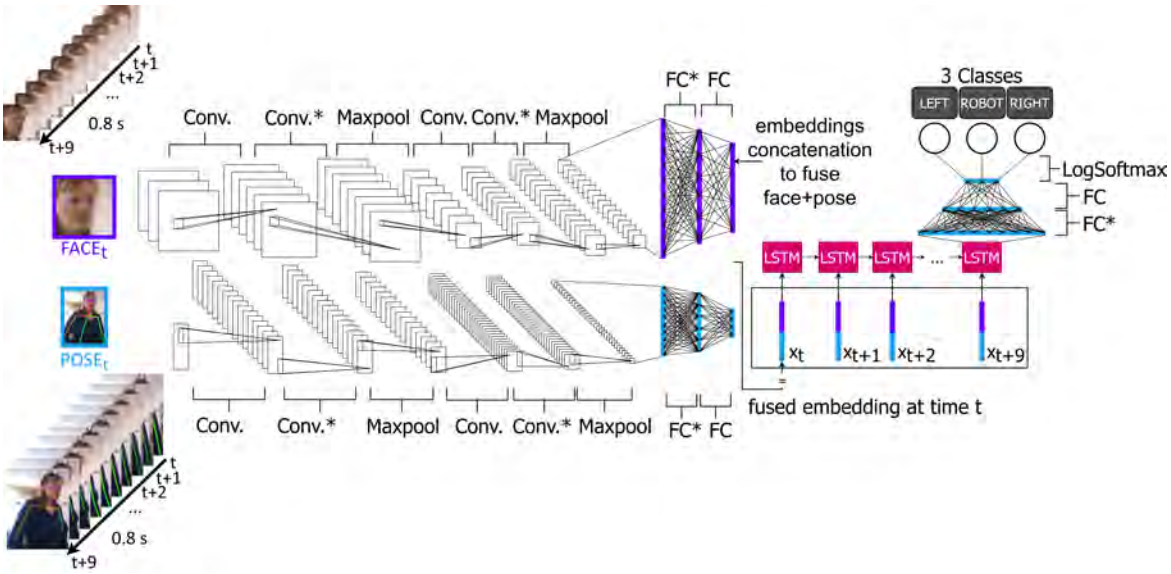

Figure 5.4 **Illustration of the Deep Neural Network for Addressee Position Estimation employing an intermediate fusion approach (Exp. 1a).** Face images and body pose vectors are passed separately to two blocks of convolution, each including two 2D convolutional and one max-pooling layers. Then, the two embeddings resulting from fully connected layers are concatenated and sequences of 10 fused embeddings are passed to the LSTM layer. The output is provided after two others fully connected layers and a LogSoftMax layer. * represents LeakyReLU activation function. See Table 5.3 in Section 5.6 for full details.

output are concatenated before the LSTM layer. In this way, features are extracted separately in convolution and then combined at a higher level of abstraction. This was inspired by a gradual fusion of modalities at an intermediate level of the network, which has been demonstrated beneficial [176, 208], and by the training on joint representations of temporal sequences as, for instance, in Nakisa et al. [159], which proved fusing streams between the convolutional and LSTM layers being beneficial rather than a late fusion after temporal training.

Therefore, the APE model employing the intermediate-fusion approach (Exp. 1.a) consists of two blocks, each including two 2D convolutional layers (the second followed by a LeakyReLU activation function) and one max-pooling layer. The two blocks for convolutions are followed by two fully connected layers (the first followed by a LeakyReLU activation function) providing the embeddings of the modalities to be concatenated. Up to this point, the 10 face images and body poses of each sequence are carried out in parallel. The fusion of the two streams was carried out as a simple concatenation, with the body pose embeddings repeated 29 times so as to balance information in the final embedding. The sequence of 10 fused embeddings is then passed through the LSTM layer. Eventually, after two final fully

connected layers (the first followed by a LeakyReLU activation function), the output is given by a LogSoftMax layer. For an illustration of the architecture see Figure 5.4, whereas for full details see Table 5.3 in Section 5.6.

Several variants of the above-mentioned intermediate-fusion model (Exp. 1.a) have been designed for additional experiments. To compare it with a late-fusion approach, a second model was trained using both visual modalities but combining them after the LSTM layer (Exp. 1.b, see Table 5.4). Also, to test performances using single modalities, mono-stream models were trained: one related to the face images (Exp. 1.c), the other for the body pose vectors (Exp. 1.d). In these two latter cases, the convolutional part of the two mono-stream models matched the two-stream one for that modality but differed in the LSTM layer, which was designed to be trained on single modalities (see Table 5.5).

As a final additional variant (Exp. 1.e), the same architecture of the intermediate-fusion model was trained to classify a binary output instead of the three-class one. The model was designed to answer an additional, parallel question: *if* the robot was addressed by the addresser. Hence, with respect to the previous task, this classification did not consider the position of the addressee in case the robot was not addressed by the addresser. For this reason, data referred to labels "LEFT" and "RIGHT" were combined for the negative answer "NOT-ADDRESSED", whereas data referred to the label "GROUP" were added to the "ROBOT" ones for the answer "ADDRESSED".

5.2.1.4 Training Procedure

To train and test the model, 10-fold cross-validation was established. In this way, the prediction of the classes could be evaluated based on the average performance of the model when trained and tested on different sets of data. To create the 10 different train and test sets, the dataset of sequences derived from the pre-processing of the Vernissage corpus was partitioned along the 10 multi-party interactions of the original corpus. Each interaction comprised two agents for a total of 20 speakers/addressers. The ratio to create the train sets was 9:10, with the remaining 1:10 for the test set. Accordingly, each train set included all the face image and body pose sequences of 18 participants, whereas the test set the ones of the remaining 2 participants. From the train sets, 90 sequences (30 for each class) were randomly extracted, removed, and used for the validation phase during the training in order to check the trend of the loss function.

Following the pre-processing pipeline (see Section 5.2.1.2), the Addressee Position Estimation model was thus trained 10 times, one for each train set, and evaluated on as many test sets. The model was fed with temporal sequences of data in mini-batches of 10 sequences.

Each sequence included 10 face images and 10 body pose vectors. The convolutional section of the network for face images was trained by employing Stochastic Gradient Descent (SGD) optimizer whereas the one for body pose vectors and the LSTM section were trained using the Adam algorithm [119]. Cross entropy was used as the criterion to compute the loss function. The model was trained for 50 epochs with a learning rate of 1e-3, with a decay of 0.1 (multiplicative factor) after 40 epochs. To prevent overfitting, a method of early stopping was implemented to stop the training after 10 trials in which the loss function of the evaluation phase increased. The model was implemented using PyTorch 1.12 (Python version 3.8), whereas the training was carried out through an NVIDIA Quadro RTX 5000 with 16 GB of RAM.

5.2.1.5 Evaluation Metrics

The classical metrics to evaluate binary classification tasks would not be suitable to measure the performance of the three-class classification model. Accordingly, Precision, Recall, and F1-score are estimated for each class and expressed as a percentage. Hence, for each class, Precision is computed as the ratio between the correctly predicted labels and the number of positively predicted labels. Recall is computed as the ratio between the correctly predicted labels and the number of actual positive labels. F1-score is the harmonic mean of Precision and Recall. Results for each class are subsequently weighted for the number of samples of each class and averaged to provide a performance of each model in terms of Weighted Precision, Recall, and F1-score. Eventually, the results of the 10 testing from the 10-fold cross-validation were averaged to obtain a final estimate of the model’s performances.

The training and the first testing of the model were achieved by keeping the 10-frames sequences as data chunks. However, since sequences were extracted from utterances, the same metrics could be used to verify the model’s performance in predicting the addressee of an entire utterance. The utterance classification was computed by averaging the predictions of all the sequences belonging to that utterance, weighted for the prediction score provided by the LogSoftMax layer.

For the binary classification task, Precision, Recall (Sensitivity), F1-score, and Specificity were calculated in the following way and expressed as a percentage. Considering the positive prediction as “the robot is addressed” and the negative one as “the robot is not addressed”, Precision is computed as the ratio between the True Positive (correctly predicted labels) and the total number of positively predicted samples. Recall is computed as the ratio between True Positive and the total number of actual positive samples. F1-score is the harmonic mean of the Precision and Recall, whereas Specificity (true negative rate) is computed as the ratio

between the correct negative samples and the total number of actual negative samples. In addition to these parameters, for a further measure of the model’s performance, an overall-F1-score of the two classes ("ADDRESSED" vs "NOT-ADDRESSED") was computed as in the three-class model.

5.2.2 Results

5.2.2.1 Performance of the model on the Vernissage Dataset

The model Addressee Position Estimation (APE) has been trained for a three-class classification task to predict the position of the addressee based on the non-verbal behavior of the addresser. To train and test the model, two visual modalities were used: face images and body pose vectors of the addresser, ordered in temporal sequences of 10 frames. Several experiments have been conducted varying the design of the model:

- Exp. 1.a: two-modalities combined with an intermediate-fusion approach
- Exp. 1.b: two-modalities combined with a late-fusion approach
- Exp. 1.c: single modality: the model was trained with only face images
- Exp. 1.d: single modality: the model was trained with only body pose vectors
- Exp. 1.e: the same model of Exp. 1.a was trained for a binary classification.

In Exp. 1.a the average performance of the intermediate-fusion model in terms of weighted F1-score was 75.01% (see Table 5.2 and Figure 5.5). This performance resulted from testing the model on single sequences without combining them in utterances (Figure 5.3 illustrates the difference between sequences and utterances). Each utterance could comprise several sequences. Hence, when considering sequences of the same utterance together, the performance increased up to 76.48%. An additional score was computed only focusing on the first sequence of each utterance, which means measuring the model’s performance in providing a correct prediction at 0.8 sec from the beginning of the utterance. Considering the limited amount of time, the model performed satisfactorily, with an F1-score of 74.15%.

For what concerns Exp. 1b (late fusion), the model reached an average F1-score of 73.18% on the sequences, 74.19% considering sequences combined in utterances, and 71.88% on the first 0.8 sec of each utterance (see Figure 5.5). Therefore, although the difference between the performances of the two approaches is not substantial, the intermediate-fusion model achieved greater performances. Respectively, the only-face and only-pose models achieved

Table 5.2 **Performances of the Addressee Position Estimation model.** Results of the 10-fold cross-validation experiments (Exp. 1.a-b-c-d) are provided in terms of mean and standard deviation of F1-score, Precision, and Recall. Performances are computed considering each sequence separately (0.8 sec)

	F1 score		Precision		Recall	
Model	avg	std	avg	std	avg	std
Exp. 1.a Intermediate Fusion	75.01	8.60	77.68	7.71	75.08	8.18
Exp. 1.b Late Fusion	73.17	7.57	76.27	6.13	73.60	6.88
Exp. 1.c Face	72.83	5.86	76.50	6.31	72.95	5.37
Exp. 1.d Body pose	72.60	6.75	75.04	6.41	72.42	6.48

an average F1-score of 72.83% and 72.61% on sequences, 73.22% and 71.05% on the whole utterances, and 72.07% and 70.77% on the first sequence of each utterance.

Weighted Precision, Recall, and F1-score have been computed for each class ("LEFT", "ROBOT", "RIGHT") to observe if the performances of the four models (intermediate-fusion, late-fusion, only-face, only-pose) were equally distributed among classes. As reported in Figure 5.6, with the exception of the only-pose, the other models present higher performances for the "LEFT" and "RIGHT" classes. For instance, in the only-face and late-fusion models, "LEFT" and "RIGHT" overcome the "ROBOT" by more than 15%. Such a difference is slightly mitigated in the intermediate-fusion model, which reflects the influence of the body pose features. Albeit lower in "LEFT" and "RIGHT" than other models, in the only-pose model, the average F1-score is similar in the three classes, measuring 72.95%, 68.56%, and 68.87% for "LEFT", "ROBOT", and "RIGHT", respectively. Such a result is due in particular to a greater recall for the "ROBOT", which exceeds 75% and overcomes precision of about 10%, meaning that this model is more permissive for the “self”.

The present models are trained and tested on sequences of data lasting 0.8 sec. In this way, for longer utterances, multiple classifications are available so that fresh update predictions are released every 0.8 sec. Figure 5.7 displays the F1-score of the four models for incremental time intervals, showing how the models perform in predicting the addressee of utterances as time passes. For instance, if at 0.8 sec the intermediate-fusion model achieves an F1-score of 74.15%, the performance increases up to 76.48% at 1.6 sec, maintains 76.5% at 2.4, and still improves up to 79.8% after 2.4 sec. Except for the only-pose model, whose performances measure about 70% and do not improve over time, the only-image and the late-fusion models follow the same trend as the intermediate fusion one, achieving an F1-score of 78.25% and 77.2% after 2.4 sec respectively.

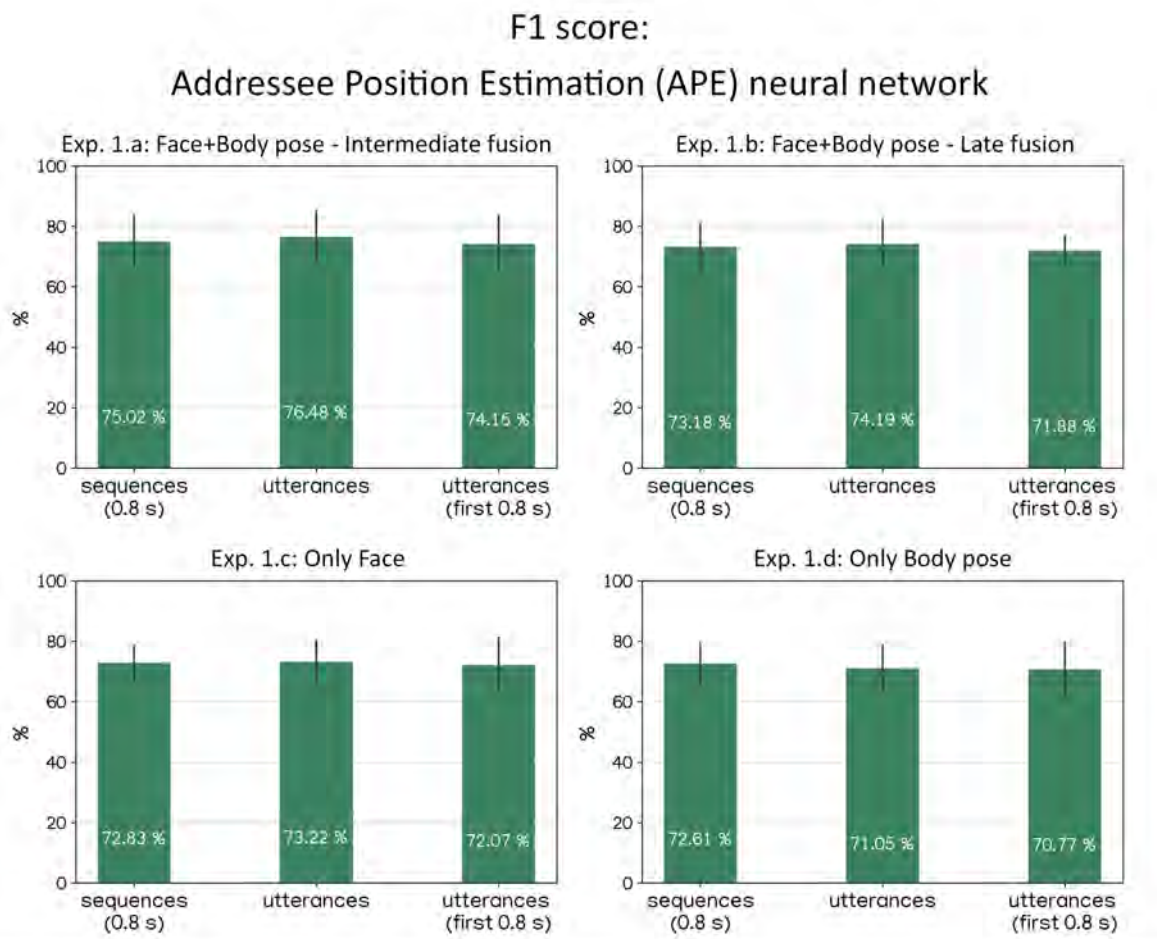

Figure 5.5 **Bar plots reporting performances of the Addressee Position Estimation model.** Results of the 10-fold cross-validation experiments (Exp. 1.a-b-c-d) are provided in terms of mean and standard deviation (error bar) of F1-scores. Performances are computed in three ways: considering one prediction for each sequence separately (0.8 sec), considering one prediction for each utterance, and considering the prediction of the first sequence of each utterance (first 0.8 sec of each utterance). On the y-axis the performance score is expressed in %.

State-of-the-art models on the Vernissage dataset were implemented as binary classification, predicting whether either the robot or another user was the addressee of an utterance. To compare our approach with the ones used by Sheikhi et al. [200], in the Exp. 1.e a model was trained to solve the following binary classification task: if the speaker was addressing the robot. The model was designed and trained as the intermediate-fusion model of Exp. 1.a (described in Section 5.2.1.3) but the last layer was modified to give a binary output. Figure 5.8 displays the results of the testing phase in a 10-folds cross-validation, where Precision, Recall, F1-score, Specificity, and overall-F1-score of the model are calculated as described

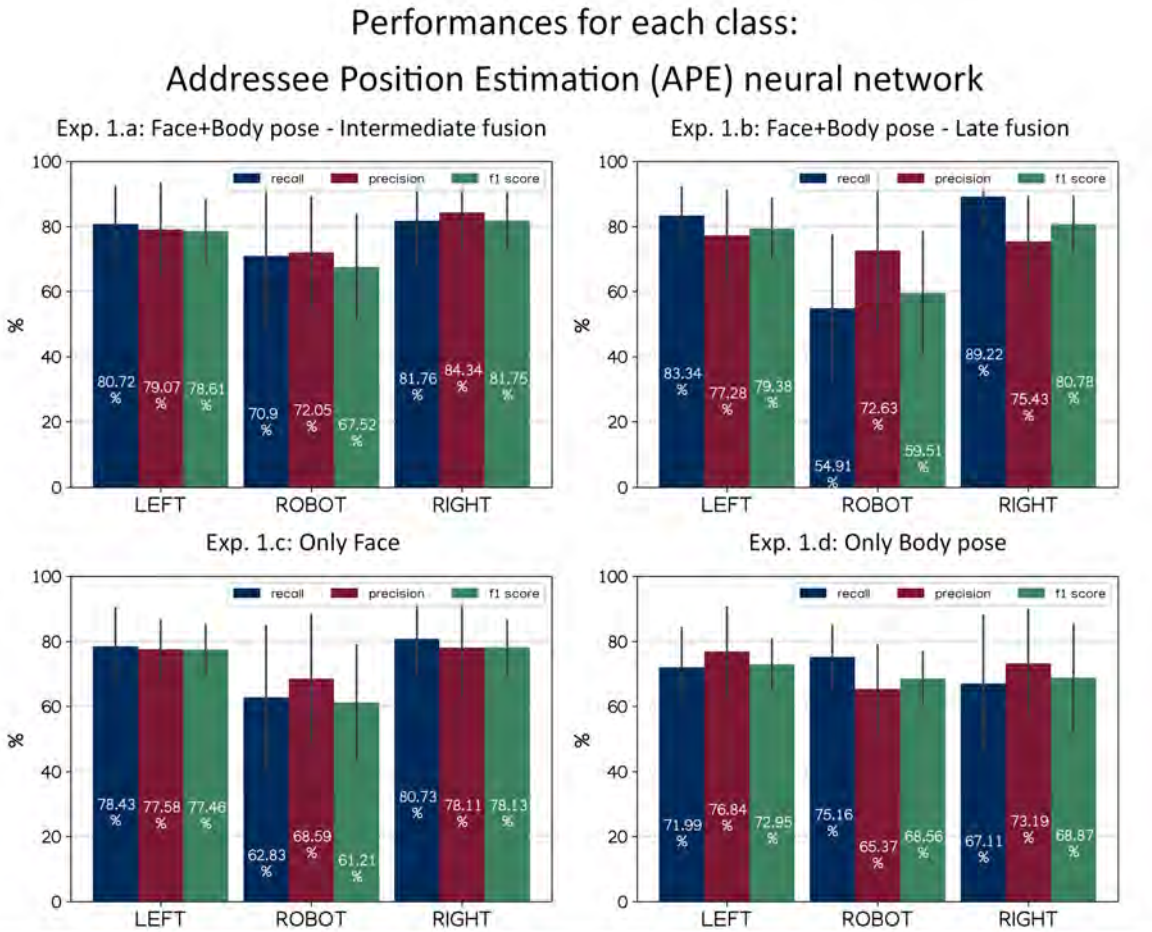

Figure 5.6 **Bar plots reporting performances of the Addressee Position Estimation model for each class.** Results of the 10-fold cross-validation experiments (Exp. 1.a-b-c-d) are provided in terms of mean and standard deviation (error bar) of Recall, Precision and F1-score for each of the 3 classes. On the y-axis the performance score is expressed in %.

in 5.2.1.5. Asking the model to predict whether the speaker was addressing the robot, considering single sequences, the average model's Recall to the affirmative answer was 73.78%, whereas the average Precision was 74.23%, and the average F1-score was 72.73%. Sensitivity achieves 80.7%. The general performance of the model, as measured by overall-F1-score was 77.36% if measured on single sequences, 79.7% considering the whole utterances, and 79.97% with respect to the first 0.8 sec of each utterance. The performance of the model in Sheikhi et al. [200] was 76.3% utterances correctly predicted employing a measure of the speaker's visual focus of attention automatically computed as input. Therefore, our model slightly outperforms the state-of-the-art. In addition, our methodology allows achieving such results before the end of the utterance: just 0.8 sec from its beginning. Moreover, the inputs

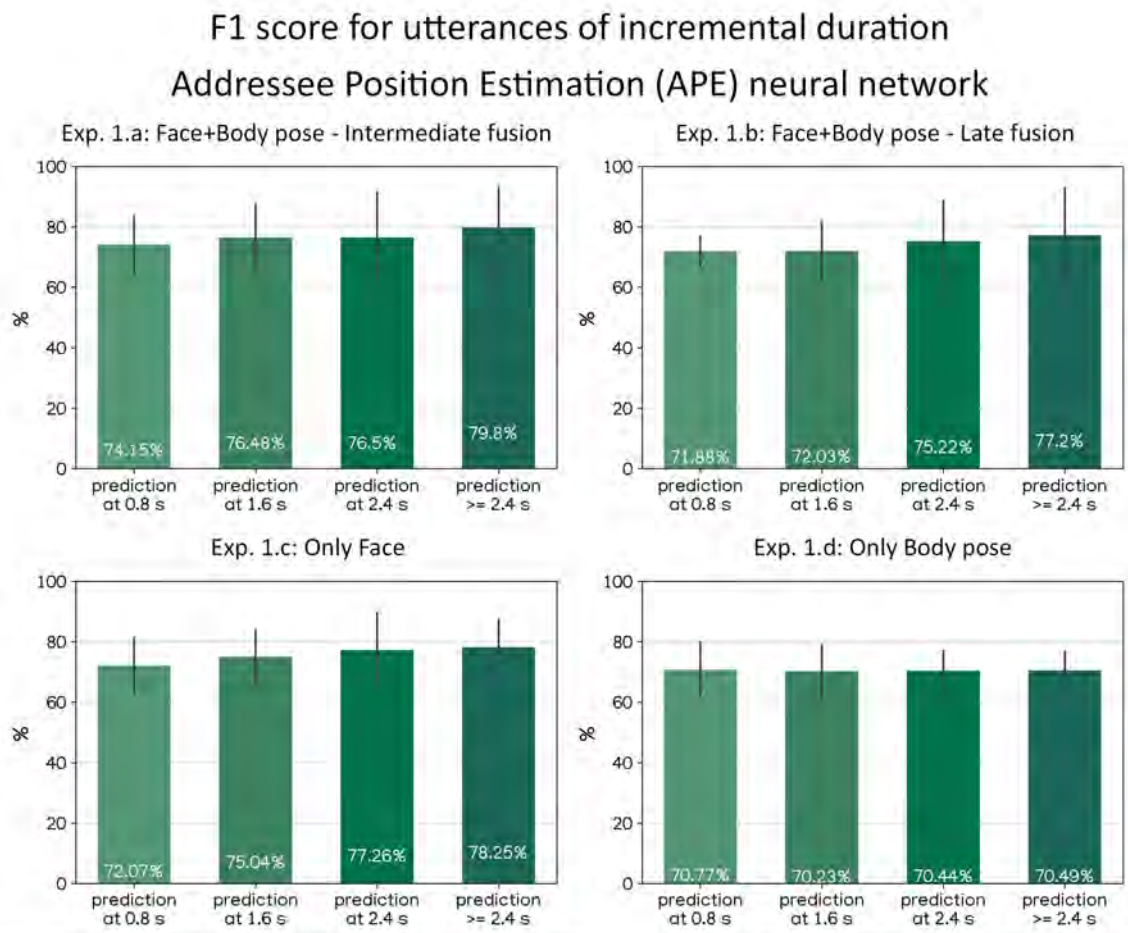

Figure 5.7 **Bar plots reporting performances of the Addressee Position Estimation model as a function of the duration of the utterance.** Results of the 10-fold cross-validation experiments (Exp. 1.a-b-c-d) are provided in terms of mean and standard deviation (error bar) of F1-score according to the duration of the utterance. Performance are computed considering the first 0.8, 1.6, and 2.4 sec. of each utterance and for utterance lasting 2.4 sec or more. On the y-axis the performance score is expressed in %.

to predict the addressee do not involve any contextual information such as the length of the utterance and the possible targets of attention in the environment as, for instance, in Sheikhi et al. [200]. Prediction is just obtained through the automatic extraction of the addresser's face and body pose from the visual stream of the robot's cameras.

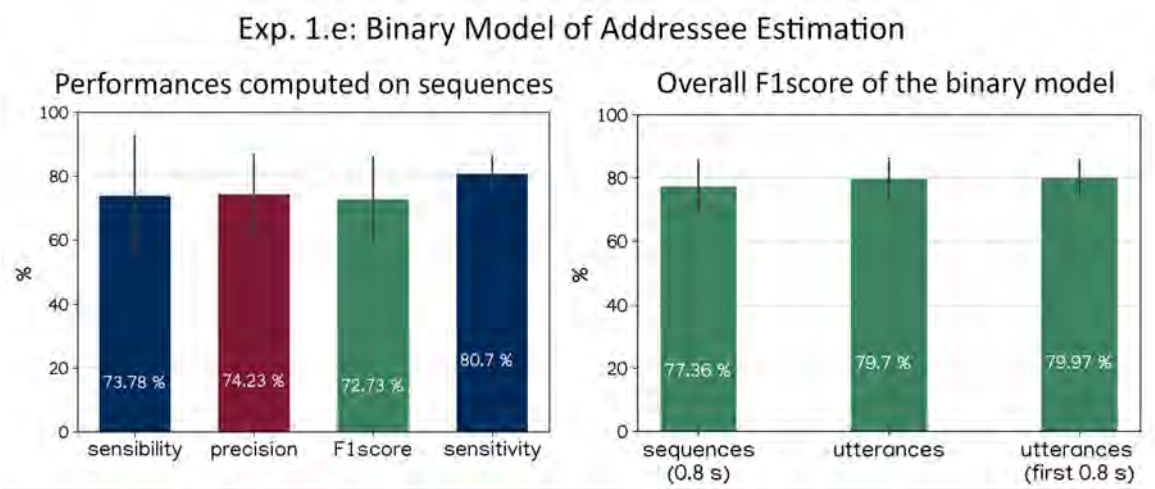

Figure 5.8 **Bar plots reporting performances of the binary classification model.** Results of the 10-fold cross-validation experiment (Exp. 1.e) are provided in terms of mean and standard deviation (error bar). Sensibility, Precision, F1-score, and Sensitivity on sequences of 0.8 sec are reported on the left. On the right, performances in terms of overall-F1-score are computed in three ways: : considering one prediction for each sequence separately (0.8 sec), considering one prediction for each utterance, and considering the prediction of the first sequence of each utterance (first 0.8 sec of each utterance). On the y-axis the performance score is expressed in %.

5.2.3 Discussion

5.2.3.1 Performance of the model

In this work, the problem of Addressee Estimation has been divided into two sub-tasks: a prediction of the position of the addressee (Addressee Position Estimation – APE model) and a final identification of the addressee by triangulating information from the addresser, the environment, and the self (Triangulated Addressee Identification – TAI module). Within this context, Addressee Position Estimation has been conceived as a three-class classification task based on the non-verbal behavior of the addresser as a cue to predict the position of the addressee with respect to the addresser, from the ego-centric perspective of the robot. Hence, the three possible classes are "ROBOT", "LEFT", and "RIGHT" referring to the addressee respectively being the robot, at the left or at the right of the addresser.

To solve this task, a deep-learning model with convolutional and LSTM layers has been designed and trained with sequences of face images and body pose vectors of the addresser. As Skantze [206] reported, information from the addresser’s head pose, as a proxy of visual focus of attention is highly relevant for humans when estimating others’ addressee and beneficial when implementing automatic Addressee Estimation models. Results from the

testing phase of the APE model corroborated this perspective: with only information about the addresser’s face, the model could predict the position of the addressee. Moreover, considering the fact that, in the Vernissage dataset, the speakers were often involved in triadic interactions with pictures on the walls, focusing their gaze on objects different from the addressee, the classification score may be considered a feasible result. Interestingly, the model trained with only body pose information proved to be equally effective. It is true that also the body pose vectors contained information about the head direction, but this was gathered only by 5 key points: 1 for the Nose, 2 for the Eyes, and 2 for the Ears. Compared to the face image, the body pose presented evidence of the speaker’s whole-body direction. Though the difference was not substantial, the models trained with both visual modalities (face and body pose) performed better than the single-modality ones. This outcome was expected, as well as the fact that the intermediate-fusion approach resulted to be more effective than the late-fusion, as literature on the topic suggested [176, 208].

The beneficial effect of combining the two visual features may be explained by analyzing the performance of each class more thoroughly. Although the overall performance of the two single-feature models was nearly identical, relevant differences appear when considering each class separately. As it appears in Figure 5.6, the only-face model predicts with higher F1-score the "LEFT" and "RIGHT" classes with respect to the "ROBOT" with a gap greater than 15%. Different is the case for the only-pose model, whose performance is more stable along the three classes. What impacts more in such a result is the high recall for the "ROBOT", meaning that the model recognizes the "ROBOT" more easily from body poses than face images. This situation seems to be reflected in the performance of the intermediate-fusion model, which combines a high performance for the "LEFT" and "RIGHT" classes with results more balanced for the "ROBOT". This suggests that beyond a general increase in performance given by the help of two channels instead of one, relevant features for "LEFT" and "RIGHT" are provided by the face modality, whereas for "SELF" by the body pose of the speaker. Interestingly, this pattern is not shared with the late-fusion model, in which the same gap between "LEFT"/"RIGHT" and "SELF" is even more evident than in the only-face model. Accordingly, this difference between the intermediate and the late-fusion approach suggests that fully-connected layers are not enough to optimally balance the two modalities.

The APE model has been developed so as to have predictions independent from the utterance length and available less than 1 second after the utterance start. The method chosen to solve this task has been to focus on sequences of data lasting 0.8 sec. For each utterance, this allowed providing a first prediction about the addressee at 0.8 sec., as well as other continual predictions every 0.8 sec., incrementally weighted on all the predictions of

sequences of that utterance. Longer utterances are formed by a higher number of sequences, hence more ample evidence for a correct estimate. This is what appears from results in Figure 5.7. The longer the addresser talks, the better the estimate of the addressee. The intermediate-fusion, late-fusion, and only-image models share this pattern. Conversely, this is not the case for the only-pose model. It appears, therefore, that the two multi-modal models inherit this characteristic from the only-face model. Moreover, one may speculate that the reason underlying this different behavior is that humans often turn their heads while speaking, in particular, if they are referring to some objects in the environment, as is the case in triadic interactions. The situation might be different for body poses that, although including information about the face pose, are more stable if considering the whole body, at least in the scenario of the Vernissage Corpus.

5.2.3.2 Insights on the three principles for the model design: Awareness of bodily non-verbal behavior, Temporality, Suitability for ecological scenarios

The model was designed to follow three general principles: awareness of bodily non-verbal behavior, temporality, and suitability for ecological scenarios. More specifically, the *awareness of bodily non-verbal behavior* has been implemented as exploiting visual non-verbal information from the addresser’s body (face and body pose) to estimate the addressee of an utterance. Such an embodied solution for the task highlights how important non-verbal behavior may be to correctly interpret verbal information and advocates making further use of this component for developing robots as conversational agents. Bodily non-verbal behavior offers profound insights into other agents’ intentions. Targeting bodily non-verbal behavior is a valid solution to improve robot’s skills of human awareness: an ability that is crucial to enhance natural human-robot interaction.

The principle of *temporality* influenced the design of the classification task and of the neural network. Firstly, temporality was applied *within each utterance*, because utterances were not considered as a whole within the neural network but were partitioned in multiple time intervals of 0.8 sec., each of them generating a prediction, as explained before. Secondarily, temporality was conceived *within each sequence*, i.e., time interval of 0.8 sec, because they were not considered as snapshots but were themselves temporal sequences consisting of 10 frames.

In future work, the principle of temporality could be also applied to enhance Addressee Estimation *within each dialogue*. Each dialogue comprises multiple utterances exchanged by at least two agents. Previous literature highlighted the importance of this kind of information,

such as previous speakers and/or previous addressees, as additional sources for Addressee Estimation [206]. For the moment, the present model does not envisage this contextual knowledge which, however, could be investigated when implementing the second module of the architecture: the Triangulated Addressee Identification.

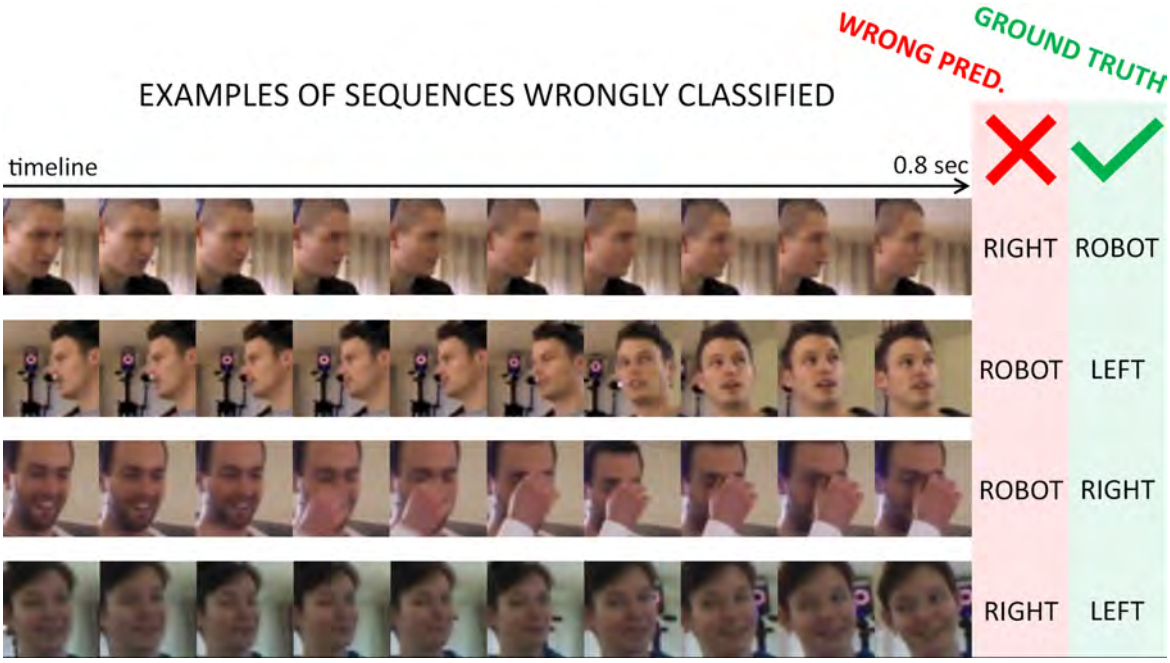

Figure 5.9 **Examples of sequences wrongly predicted.** The face images of four sequences are exhibited reporting the wrong prediction given by the intermediate-fusion model (Exp. 1.a) and the ground truth (correct addressee).

As a third principle, the *suitability for ecological scenarios* drove the selection of the dataset [107]. The APE model has been designed and trained to be implemented on the iCub robot as a future step. Hence the importance of relying on data acquired directly from the robot’s cameras (although the robot used for the Vernissage dataset recording was a Nao). Having a dataset recorded in embodied interaction with a physical robot is indeed essential to train the model to be tested in embodied physical HRI. People behave differently in front of a physical robot, or a virtual agent [133].

Another element following the ecological scenario principle was related to triadic interactions. Instead of having a fixed conversational scenario, e.g., everybody sat at a table directly looking at their addressee, the Vernissage corpus envisages situations more difficult to predict. For instance, Figure 5.9 shows some sequences wrongly predicted. It appears that the head direction was not always predictive of the addressee’s position and that in certain cases it even causes errors. Triadic interactions are one of the causes of such errors because in the

dataset participants could talk to Nao while looking at a picture on the right or, vice versa, talking to their companion on the left while looking at a picture collocated over the robot. It is true that this may result in lower performance of the model and in higher confusion of features belonging to different classes. However, this confusion has been balanced with the temporal design of the model that is aimed at considering frames, sequences, and predictions varying over time and is more representative of real-time interactions.

5.2.3.3 Limitations and future work

The current APE model presents three major limitations. Even if it is built upon the principle of ecological validity, the model does not envisage the presence of a large number of people in the environment. It is true that with respect to most of the previous implementations of Addressee Estimation in HRI, this model does not merely provide a binary outcome. Nevertheless, a three-class classification of the position of the addressee might not be enough in case of crowded places such as airports, hotel reception areas, malls. Relying on auditory cues might not represent a solution, since a crowded environment often involves noisy auditory information. Conversely, non-verbal bodily behavior still allows for more discretized spatial representations in the longitudinal dimension (adding for instance also extreme left and extreme right positions) or in depth.

As a second limitation, the ape Model does not provide a final estimate of who the addressee is: except for the class “ROBOT”, the model only predicts the addressee’s position with respect to the addresser. However, future work will overcome this restraint. Indeed, the outcome of the APE model represents only the first core step for an Addressee Estimation architecture inspired by Shared Perception. As anticipated in Section 5.1.3, a module of Triangulated Addressee Identification (TAI) will take as input the prediction of the APE model and triangulate this information with the one from the environment and the self, achieving, the identification of the addressee.

Eventually, for the moment the model has been designed only for visual information. No auditory or contextual information is used. However, by design, the model was conceived not to rely on contextual information. In this way, a first prediction of the addressee may be provided by only focusing on visual information, without any knowledge of the number of agents in the environment, previous addressees, previous speakers, topic of the dialogue, hot-words, etc. Some of this information may be used instead in the TAI module, and, hence, exploited in future implementations. Future work should also envisage the use of auditory cues, which comprise not only verbal information but also, for instance, prosody. Auditory data was already available in the Vernissage corpus. It was recorded by the Nao’s

microphones, but too noisy to extract reliable prosodic information. Therefore, prospected plans consist of recording also auditory information in a forthcoming data acquisition during HRI with the iCub robot.

5.3 Exp. 2. Preliminary assessment of generalizability of the model on iCub

The data used to train and test the model of addressee position classification were recorded in a multi-party interaction with the robot Nao, using the robot's cameras. Nevertheless, the model was thought to be subsequently implemented on the iCub platform. The future plan consists of employing a transfer learning approach to retrain the model on data taken directly in interactions with the iCub robot and make it effective on another platform. However, we thought that testing the model on already available data recorded by the iCub's camera might provide a preliminary assessment of its robustness and generalizability.

5.3.1 Methods: dataset and procedure

To this aim, a corpus of data that had already been collected was exploited. Data consisted of camera frames recorded from 10 participants standing in front of the iCub robot and pretending to talk to an imaginary addressee positioned in three different positions. The addressee's positions were the same as the Vernissage corpus: "LEFT" and "RIGHT" of the participant from the robot's perspective and "ROBOT", meaning that the robot was the addressee. The iCub's corpus had been labeled by the experimenters directly during the recording. The recording session did not involve reciprocal interaction with the robot, which only played the role of observer. During the session, participants stood in different positions in the room at a distance of about 1.8 m from the robot. They had to draw the robot's attention, endowed with a model for autonomous sound localization and face tracking [83]. Once the robot turned its face and looked at the participant, the experimenter asked the participant to turn and look towards three different directions: "LEFT", "RIGHT", or "ROBOT". For each pose, the view from the robot's left camera was recorded for 5 sec.

Though this dataset was similar to the Vernissage Corpus in terms of labeling and addressee position, noticeable differences occur. The first main difference between the two datasets is that participants interacting with the iCub wore facial masks because data was recorded during COVID-19 restrictions. Furthermore, in the Vernissage corpus, participants behaved and moved naturally while speaking and often referred to objects in the environment,

which provided dynamism to data organized in temporal sequences. Conversely, in the iCub dataset, participants just hold the same position throughout the 5-sec recording, which removed any dynamism from the data. Therefore, since the data within each recording did not differ from each other, to test the model on the iCub dataset, only the first sequence of 10 frames for each recording was considered. Body pose vectors and face images were then extracted from the frames with the same methods used for the Vernissage corpus (see Section 5.2.1.1) leading to a set of 30 sequences for each label, each consisting of 10 body pose vectors and 10 face images.

5.3.2 Results

After a final training on the whole dataset, the Addressee Position Estimation model was tested on data recorded from an iCub robot's cameras, as specified in Section 5.3.1, in four experiments: (1.a: intermediate-fusion, 1.b: late-fusion, 1.c: only-face, 1.d: only-pose). Figure 5.10 shows the four confusion matrices that report the Recall and Precision for each class and the general F1-score of the model. Results are only related to the testing on sequences because the composition of the dataset did not include any information about utterances.

The intermediate-fusion model, which on the Vernissage dataset outperformed the others, achieved an F1-score of 65.14%, whereas the late-fusion model reached 58.47%. The two single-modality models, which produced similar results on the Vernissage dataset, dramatically differ in this case. The model trained by using only body poses did not even reach 50%. Conversely, the only-face model outperformed the intermediate-fusion multi-modal one and obtained an F1-score of 69.25%.

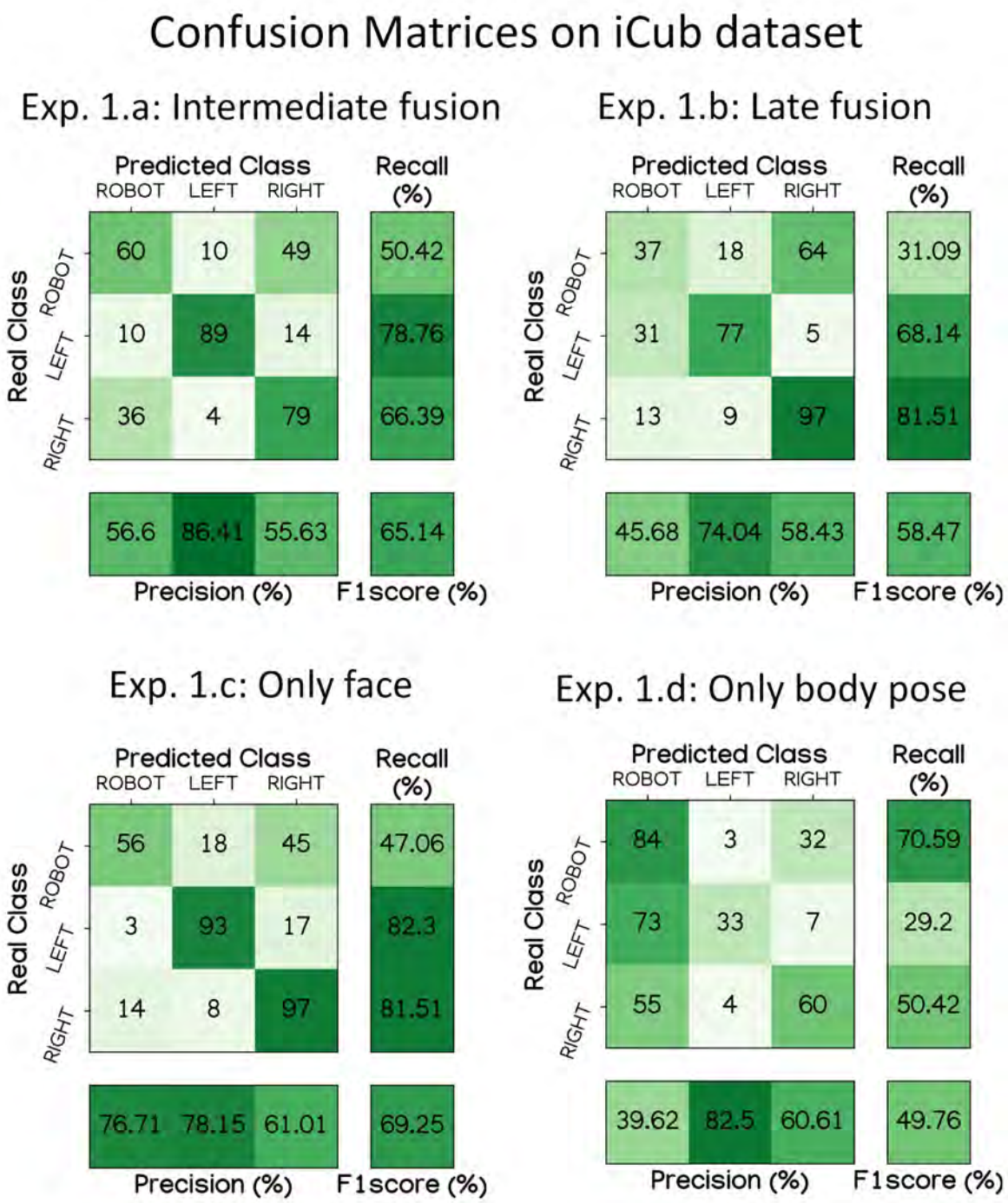

Figure 5.10 **Confusion Matrices.** Generalizability performances of the APE model on the dataset recorded by iCub in 4 experiments (1.a-b-c-d). Within the 3x3 matrix values represent the number of sequences, whereas Recall, Precision, and F1-score are expressed in %.

5.3.2.1 Discussion

A design following the principle of suitability to ecological scenarios should hypothetically result in a higher level of portability between robots. As mentioned above, the APE model was designed to be implemented on the iCub robot. To achieve this goal, one of the next steps for future work will be acquiring some data with the iCub and use them following a transfer learning approach to enhance the model performance. The portability of models from one robot to another might result in a performance loss. In this particular case, this could be due to the different heights where the two robots' cameras are located and, accordingly, to their different perspective on the environment.

A small dataset recorded through the iCub cameras and labeled with information about the (pretended/imaginary) addressee was already available (see Section 5.3.1). Hence, this dataset has been used for a preliminary check of the model's portability. After training the four models (intermediate-fusion, late-fusion, only-face, only-pose) on the entire Vernissage corpus, such models have been tested on the iCub dataset. Between the two mono-modality models, the only-face performed better, with an F1-score slightly lower than with the Vernissage corpus. Conversely, the only-pose model suffered a considerable loss. The intermediate-fusion model was affected by this loss resulting in performance slightly worse than the only-face although still acceptable.

These results should indeed be read considering the differences between the two datasets. Beyond the different robots, the dataset recorded with iCub involved participants wearing a face mask, not really speaking with someone but only looking in the direction of an imaginary addressee. The dataset was recorded so as to train a CNN to classify snapshots of head directions. For this reason, participants were not requested to move during the recording but only to look at specific directions, so the frames in sequences were very similar to each other. It is not surprising therefore that the only-face model performed in line with previous results, whereas worsening occurs where also the pose was involved. Given all these considerations, this preliminary check suggests the model may be ported to a different platform, such as the iCub robot, with possible adjustments using transfer learning.

5.4 The design of the Shared Perception-Addressee Estimation Architecture

Addressee Estimation is a social ability that can be tackled in the frame of Shared Perception (see Section 2.5.2). Considering the core idea of Shared Perception (having at disposal three

sources of information - environment, self, others - and being able to integrate them), two elements were missing to implement a Shared Perception-Addressee Estimation (SP-AE) architecture on the iCub: the information coming from the others (APE module) and the integration of the three sources of information (TAI module).

Since the APE module was needed as input of the TAI module, we decided to start developing the former. The APE module that has been developed interprets the addresser's bodily behavior in terms of addresser/robot-relative position of the addressee. Providing information about the addresser's intentional relation toward the environment, this module represents one of the three sources of information needed for Shared Perception triangulation.

The TAI module ("Triangulated Addressee Identification") was conceived based on the three sources of information of Shared Perception, its function consisting of comparing the information from:

1. the interpretation of the speaker's behavior,
2. the robot's awareness of other social agents and prior information of previous events,
3. the robot's continual perception of the environment.

To achieve the whole SP-AE architecture, a set of other modules and connections are required. Modules related to sensory inputs, attention, perception, memory, and action generation are needed to provide the SP-AE architecture with the necessary information to identify the addressee and the cognitive skills to act autonomously. Most of these modules will be imported by another architecture recently developed on the iCub robot in our department. The HRI Tutoring Framework for Long-Term Personalization and Real-Time Adaptation [19] was developed to make the iCub a good robotic tutor of Yoga movements with the ability to adapt in real-time to participants' performance and to personalize its instructions to each participant in long-term interactions. To autonomously interact with people, the robot was endowed, among others, with modules for sound localization, person recognition and tracking, pose detection, short-term spatial memory, and long-term memory to store information about known participants.However different the objective of the architecture may be, such modules will also be exploited in the SP-AE architecture.

Figure 5.11 represents the SP-AE architecture. Below a description of all modules is provided, listing those already existing on the iCub robot and/or imported from the HRI Tutoring Framework, the APE module, and the ones that still need to be implemented.

Sensory input systems The two sensory modalities required by the architecture are vision and audio.

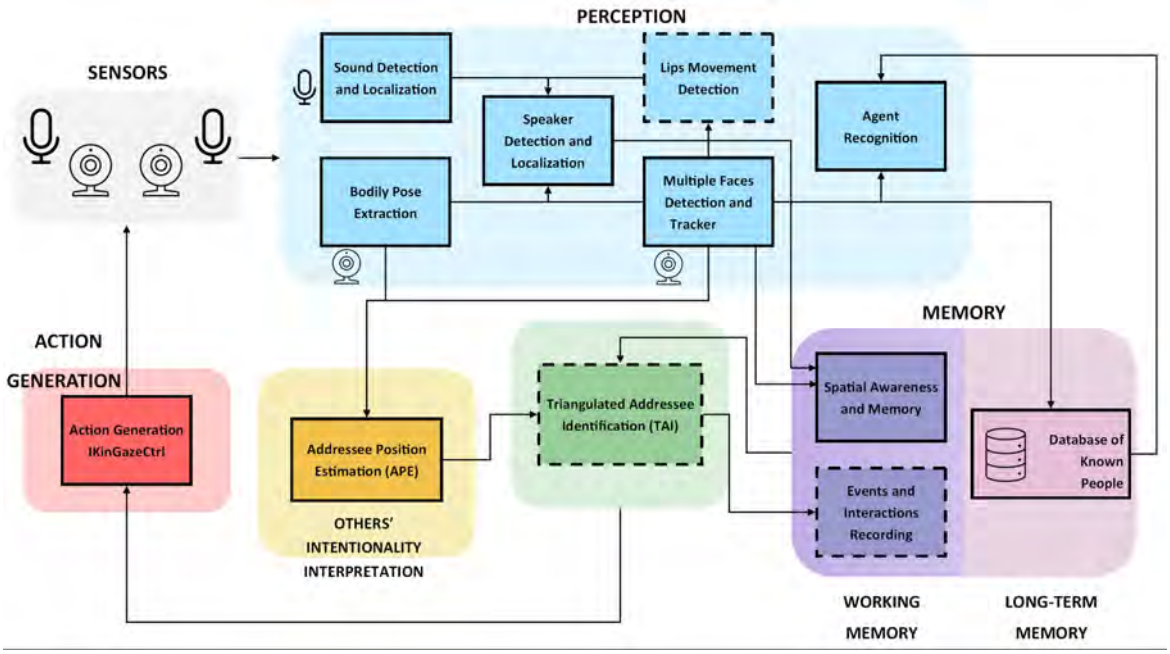

Figure 5.11 **Shared-Perception Addressee-Estimation (SP-AE) Architecture.** Modules related to perception (light blue), memory (violet), and action generation (red) contribute to the estimation of the addressee together with the APE module for the interpretation of others' intentionality (yellow) and the TAI module for triangulation (green). The modules with a continuous outline have already been developed whereas the ones with the dashed line still need to be so.

- Binaural auditory information is processed by the two microphones located in the ear cavities of the iCub head.
- Visual sensory information is provided by the two 2D cameras of the iCub platform.

Attention and Perception

- An auditory attentional system is implemented to localize sounds starting from binaural information and exploiting the auditory phenomenon of interaural time difference [83]. The same module also provides information about sound activity detection.
- A module for face detection and tracking [84] localizes faces from the visual stream of cameras and simultaneously takes records and tracks multiple faces.
- An additional visual attention system may be helpful in SP-AE architecture: a module for lips movement detection that still needs implementation should be connected to the face detection module to improve the localization of the speaker thanks to an additional sensory modality.

- A module for body pose extraction is implemented through the OpenPose algorithm [33] providing a 2D vector of 18 key points for each person detected.
- A module for Person Recognition [82] exploits embeddings previously extracted from agents' faces to recognize them in later interactions or when losing their initial tracking.

Memory

- A module for spatial working memory takes records of the number of agents and their position in the room from the ego-centric perspective of the robot through a discretized partition of the environment [19].
- A complementary working memory module keeping records over time of the speaker's identity, their addressee, and other contextual information about interaction might be required in future developments of the SP-AE architecture.
- A module of long-term memory takes records of the embeddings of agents that already interacted with the robot, to subsequently recognize them in the interactions [82].

Action Generation

- A Gaze Controller was implemented on the iCub to control the joints of the neck and eyes to track a 3D cartesian point in space [187]. This allows the robot to track objects (faces), switch attention toward new objects, and explore the environment looking for new objects.

Others' intentionality interpretation

- The Addressee Position Estimation (APE) module, as specified in section 5.2.1, predicts the position of the addressee from the ego-centric perspective of the robot. The module is implemented as deep neural network taking as input temporal sequences of face images and body poses of the speaker and predicting the position of its addressee in a three-class classification task. The predicted addressee could be positioned either at the left or at the right of the speaker or could be the robot.

Triangulation

- The Triangulated Addressee Identification (TAI) module would take in input the following information:

1. Addressee Position as predicted by the APE module, representing the information coming from the other.
2. Number of agents and their position in the environment as recorded in the spatial memory module, representing the self in terms of the robot's awareness of the environment.
3. Additional information about the position of previous and novel agents in the environment provided by visual perception while proactively exploring the environment (for instance, in case the output of the APE module does not find correspondence in the robot's awareness of agents' positions (see Figure 5.12). This represents the information coming from the environment.

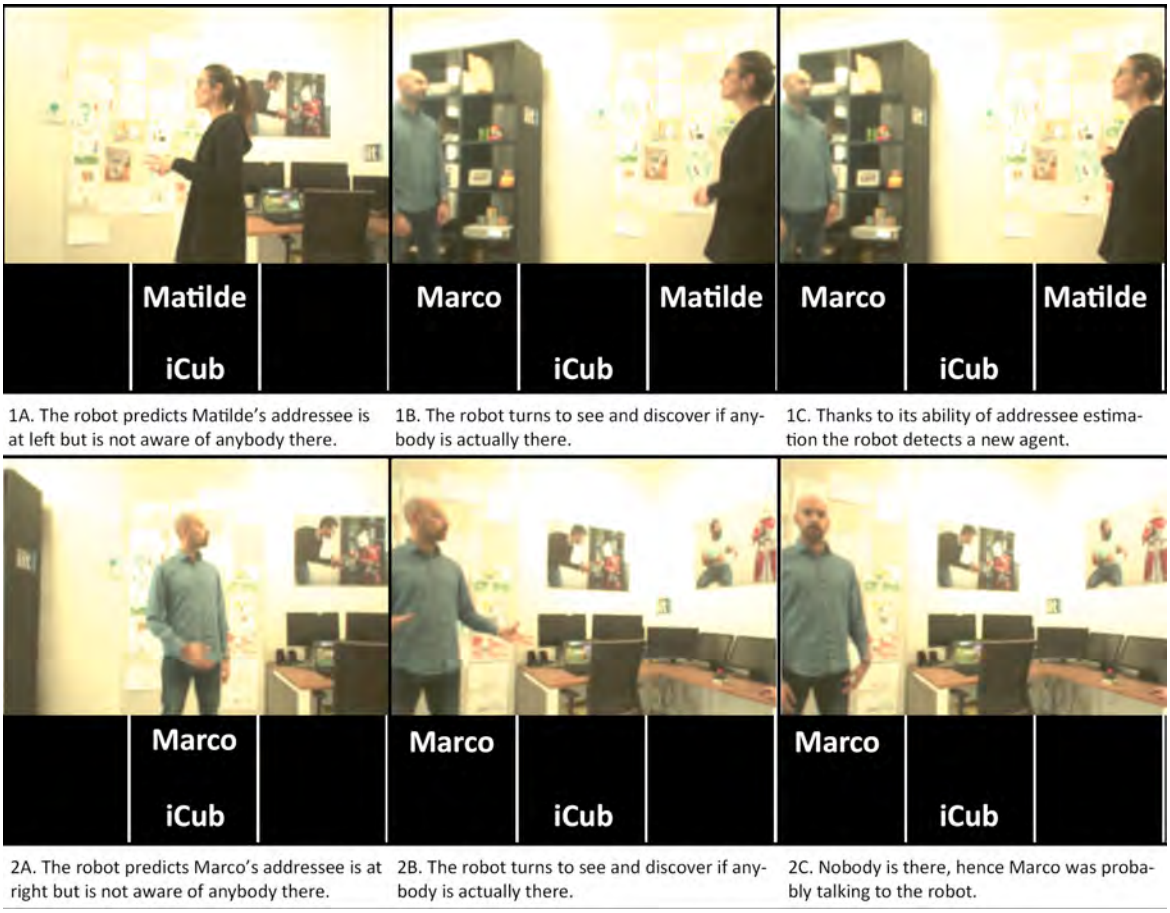

Figure 5.12 **Illustration of two possible outcomes from Triangulated Addressee Identification Module.** For each example, three snapshots of iCub's left camera taken over time are shown together with the information about the position of the known agents in the environment available in the spatial memory, and the description of the robot's behavior if it was driven by the whole SP-AE architecture.

5.5 Conclusion

The whole Shared-Perception Addressee-Estimation (SP-AE) architecture still requires several steps to be implemented on the robot. The main point is the development of the TAI module for a triangulated identification of the addressee. Here, the information coming from all the different sources should be integrated. This means that the resulting estimation of the addressee may be constantly revisited and corrected. In Chapter 2, we maintained that the Shared Perception process may bring different achievements: deepening one's understanding of others, establishing a common ground with others, and enhancing one's representation of the environment. Accordingly, solving the Addressee Estimation problem as a triangulation task may be beneficial in all these three directions.

Addressee estimation is connected to the *ability to understand others*. More specifically, it falls inherently within the HRI research quest for endowing social robots with reading intentions skills [196]. An important approach to develop human awareness in robots is focusing on human non-verbal behaviors, which often convey important information. This is also the case in Addressee Estimation. Correctly predicting whom a person is talking to is essential when a robot needs to interact with others, not only because the robot becomes able to detect whether others are talking with them, but also because in this way the robot can learn that another agent is involved in the conversation. Moreover, Addressee estimation paves the way to make the robot understand others through the interactions they keep with third agents. For instance, others might not be "angry", "excited", or "friendly" per se. Such emotions may be referred only towards third agents so that agent A could be angry with agent B, but friendly with C. Understanding others, often comes through understanding their interactions. From this perspective, Shared Perception proves to be profitable in providing a deepen and correct understanding of the social dynamics of the interaction. For instance, the triangulation of the three sources of information provided by the TAI module may result in a more precise identification of the addressee, which may differ, as in Figure 5.12, from the estimation coming from the APE module alone, since it was only based on the non-verbal behaviors of the speaker.

Understanding others' addressee enables *creating a common ground* with them, hence establishing a more natural interaction. A correct estimation of the addressee provides the robot with greater skills for effective command reception, correct turn-taking, and better understanding of the message conveyed by the speaker. Messages are often addressee-dependent indeed. Deictic expressions are words or phrases which strictly depend on the context. Several deictic expressions depend on the speaker, such as spatial deixis ("here',

"there', "this', "that'). But others rely on the addressee, such as personal deictic words ("you', "she', "he', "they'). Therefore, only a triangulated Addressee Estimation grants a correct interpretation of deictic sentences. Moreover, going back to the above-mentioned example of Addressee Estimation for emotion recognition (agent A that is angry with B but friendly with C), one may notice that correctly understanding the social/emotional dynamics among agents is also beneficial to create a common ground with them. Hence, Addressee Estimation allows the robot to better understand the situation and behave accordingly.

Eventually, Shared Perception is not only referred to understanding others but also to achieving an *augmented perception of the external world*. Addressee estimation may enhance the robot's awareness of the position or the number of agents in the environment. From the first example illustrated in Figure 5.12 it appears that, for instance, visual and auditory systems may not be the unique modalities for the robot to perceive and discover the environment. In this sense, sociality may be considered an additional sense to experience the world. Even in the animal kingdom, animals living in groups rely on each other, for example, to detect predators. In a similar way, a robot endowed with the SP-AE architecture may rely on other people to localize agents not yet detected in the room.

Addressee Estimation is only one task in which the triangulation of the three sources of information (others, self, environment) may provide interesting results, but from this perspective, Shared Perception appears to be essential to the design and development of socially interactive robots. As for Addressee Estimation, the framework of Shared Perception may inspire the development of other social skills as well, enabling robots to interact more naturally and efficiently with humans.

5.6 Appendix to Chapter 5: Descriptive tables of neural network architecture

Table 5.3 **Description of the hybrid architecture (CNN + LSTM) employed in the intermediate-fusion approach (Exp. 1.a).**

	Face Image		Body Pose Vector	
Layers	Input	Parameters	Input	Parameters
Conv1	[100, 3, 160, 160]	k=7, s=1	[100, 1, 18, 3]	k=(3,1), s=1
Conv2 + LeakyReLU	[100, 6, 154, 154]	k=5, s=1	[100, 16, 16, 3]	k=(3,1), s=1
MaxPool1	[100, 8, 150, 150]	k=2, s=2	[100, 16, 14, 3]	k=(2,1), s=(2,1)
Conv3	[100, 8, 75, 75]	k=5, s=1	[100, 16, 7, 3]	k=(3,1), s=1
Conv4 + LeakyReLU	[100, 12, 71, 71]	k=3, s=1	[100, 32, 5, 3]	k=(3,1), s=1
MaxPool2	[100, 16, 69, 69]	k=2, s=2	[100, 32, 3, 3]	k=2, s=2
Flatten	[100, 16, 34, 34]		[100, 32, 1, 1]	
FC1 + LeakyReLU	[100, 18496]		[100, 32]	
FC2	[100, 4624]		[100, 24]	
Concatenation	[100, 578]		[100, 20]	
	Input			Parameters
LSTM	[10, 10, 1158]			h_dim=516
FC3 + LeakyReLU	[10, 256]			
FC4	[10, 128]			
LogSoftmax	[10, 3]			

Table 5.4 **Description of the hybrid architecture (CNN + LSTM) employed in the late-fusion approach (Exp. 1.b).**

	Face Image		Body Pose Vector	
Layers	Input	Parameters	Input	Parameters
Conv1	[100, 3, 160, 160]	k=7, s=1	[100, 1, 18, 3]	k=(3,1), s=1
Conv2 + LeakyReLU	[100, 6, 154, 154]	k=5, s=1	[100, 16, 16, 3]	k=(3,1), s=1
MaxPool1	[100, 8, 150, 150]	k=2, s=2	[100, 16, 14, 3]	k=(2,1), s=(2,1)
Conv3	[100, 8, 75, 75]	k=5, s=1	[100, 16, 7, 3]	k=(3,1), s=1
Conv4 + LeakyReLU	[100, 12, 71, 71]	k=3, s=1	[100, 32, 5, 3]	k=(3,1), s=1
MaxPool2	[100, 16, 69, 69]	k=2, s=2	[100, 32, 3, 3]	k=2, s=2
Flatten	[100, 16, 34, 34]		[100, 32, 1, 1]	
FC1 + LeakyReLU	[100, 18496]		[100, 32]	
FC2	[100, 4624]		[100, 24]	
LSTM	[10, 10, 578]	h_dim=512	[10, 10, 20]	h_dim=256
FC3	[10, 512]		[10, 256]	
Concatenation	[10, 128]		[10, 128]	
	Input			Parameters
FC4 + LeakyReLU	[10, 256]			
FC5	[10, 128]			
LogSoftmax	[10, 3]			

Table 5.5 **Description of the hybrid architecture (CNN + LSTM) employed in the single modality approach (Exp. 1.c-d).**

	Face Image Exp. 1.c		Body Pose Vector Exp. 1.d	
Layers	Input	Parameters	Input	Parameters
Conv1	[100, 3, 160, 160]	k=7, s=1	[100, 1, 18, 3]	k=(3,1), s=1
Conv2 + LeakyReLU	[100, 6, 154, 154]	k=5, s=1	[100, 16, 16, 3]	k=(3,1), s=1
MaxPool1	[100, 8, 150, 150]	k=2, s=2	[100, 16, 14, 3]	k=(2,1), s=(2,1)
Conv3	[100, 8, 75, 75]	k=5, s=1	[100, 16, 7, 3]	k=(3,1), s=1
Conv4 + LeakyReLU	[100, 12, 71, 71]	k=3, s=1	[100, 32, 5, 3]	k=(3,1), s=1
MaxPool2	[100, 16, 69, 69]	k=2, s=2	[100, 32, 3, 3]	k=2, s=2
Flatten	[100, 16, 34, 34]		[100, 32, 1, 1]	
FC1 + LeakyReLU	[100, 18496]		[100, 32]	
FC2	[100, 4624]		[100, 24]	
LSTM	[10, 10, 578]	h_dim=512	[10, 10, 20]	h_dim=256
FC3 + LeakyReLU	[10, 256]		[10, 256]	
FC4	[10, 128]		[10, 128]	
LogSoftmax	[10, 3]		[10, 3]	

Chapter 6

Conclusion

This research considered social HRI from the point of view of perception and was motivated by the idea that Shared Perception is crucial to enhance natural and effective interaction between humans and robots. The general question underlying the research has been "how social interaction affects perception?". Nevertheless, the diverse nature of the agents involved in HRI (i.e., humans and robots) brought me to tackle Shared Perception from different perspectives: the investigation of Shared Perception in humans with a focus on the specific mechanism of Context Dependency, the development of a Shared Perception-inspired skill in robots with a focus on the socio-perceptual ability of Addressee Estimation, and, as an overall binding element, the outline of a theoretical framework of Shared Perception. Based on empirical findings and philosophical perspectives about social perception, such a general framework serves as a reference for the experimental research about Shared Perception in Chapters 3 and 4 and for the development the Addressee Estimation model inspired by Shared Perception in Chapter 5.

As a conclusion of this work, after examining how each Research Objective has been fulfilled (see Section 6.1), a final overview will exhibit the impact of Shared Perception in the field of socio-cognitive HRI (see Section 6.2).

6.1 Achievement of Research Objectives

6.1.1 RO1: Formulating a theoretical framework of Shared Perception

As defined in Chapter 2, Shared Perception is the observer's perceptual ability to integrate different sources of information: perception of the environment, private internal models of reality,and perception of other social agents. Also, the integration dynamics have been

described as triangulation among these three sources. The idea of providing a general definition stemmed from the need for a comprehensive account of Shared Perception, valid for humans and robots.

One of the achievements of the Shared Perception account has been to frame together humans and robots. The objective of Section 2.4 was to put in parallel studies about human perceptual or attentive phenomena triggered during interaction with other humans and studies about the same effects triggered by robots. Social robots can therefore convey similar information to humans. As it was demonstrated, when interacting with robots, humans are able to implicitly understand and integrate into their perception where the robot's attention is directed, what is its perspective on the environment, what hidden information about the environment is disclosed by its actions, as well as its internal state. This is a crucial feature for natural and effective HRI. Humans, which grew up in a social world and acquired social perceptual abilities since the first months of life, are prone to integrate others' intentional relation to the environment. The fact that they do the same even in front of robots raises the possibility of designing smooth and effective interactions with robots, based on the same perceptual phenomena that support human sociality.

On the same line, a fundamental aspect covered by the framework is that different social perception mechanisms are not only envisaged separately but also conceived together. With respect to previous literature, this represents one of the contributions of this work. Therefore, integrating others' attention, perspective, actions, and inner states are envisaged as a unique extensive phenomenon meant to generally grasp other more specific mechanisms under the terms of "perception of others' intentional relation to the environment".

Empirical research on distinct mechanisms of social perception is therefore the foundation of the Shared Perception account. Nevertheless, the perspective proposed here was a frame interpreting such mechanisms together because out of laboratories, where real interactions happen, such processes often occur contemporarily and influence each other. Deepening the investigation of these and other perceptual processes separately is still crucial: the studies contained in Chapters 3 and 4 testify to it. However, to enhance the quality of HRI and bring robots outside of laboratories, a comprehensive account of Shared Perception, which was missing, was needed.

6.1.2 RO2: Investigating the mechanism of Context Dependency in human visual perception of space during social interaction

Context Dependency is a well-known perceptual phenomenon that clearly shows perception is not a mere reception of stimuli. A crucial difference occurs between what is present in the external world and what we, in fact, perceive. Perception is the entire process connecting these two poles. As explained in Chapter 2, many elements might be involved in such a process. Among them, Context Dependency reveals the influence of previous experience. Information received through senses is affected by previous experience in this respect: what we previously perceived becomes an internal model of reality that acts as priors/predictions on what we are perceiving, hence modifying sensory information.

Previous literature explored this phenomenon only in individual scenarios. In an individual context, the mechanism reveals two of the three sources of information enabling Shared Perception: the environment and the self. The sensory stimulus can be seen as information from the environment, whereas the priors – the internal model of reality – can be considered part of the information related to the self. Keeping the perspective of Shared Perception, the studies reported in Chapters 3 and 4 tried to add the third source of information: the other social agent.

For the first time, in this work, Context Dependency has been investigated in a social context. An interactive experimental setting was inspired by state-of-the-art but with the addition of a humanoid robot interacting with participants and exhibiting two different styles of behavior (mechanical or social) and employed to explore how social interaction affects Context Dependency in visual perception of space. Interestingly, the robot, which had the role of stimuli demonstrator indicating visual stimuli to participants, impacted differently on the participants' perception, according to its social or mechanical behavior. With the social robot, the influence of priors was significantly lower than with the mechanical one. Both the psychophysics and computational approaches agreed on this, showing that sociality caused a rebalance between sensory information and prior information. Also, from the Bayesian modeling, sociality appeared as an additional factor, not envisaged by models validated in individual scenarios.

This phenomenon appears in line with the Shared Perception framework. In front of a social agent, perception is triangulated among information coming from the environment (sensory stimuli), information coming from the self (priors as the internal model of reality), and information coming from others (perception of the social robot). This was not the case of the robot behaving mechanically, as suggested by the results from the computational

model of Context Dependency. Considering the influence caused by the social robot, further consideration is needed on what element of the social robot affected Context Dependency. Was the influence triggered by the robot's attention (perceived through its gaze and head movements), by its action (which albeit identical was perceived as endowed with more agency), or by the social context (created by the robot's speech and emotions expression)?

For the moment, there is no element to settle the question. As it was the first time Context Dependency was investigated during the interaction, the study was designed to maximize the influence caused by the social interaction. Therefore, it might be the above-mentioned features (attention, action, and social context) contributed to the effect. This might be seen as a limitation of the study, which for the moment does not allow to outline which element (or which more) was integrated into the Shared Perception process. Future studies should move in this direction.

6.1.3 RO3: Developing a model for Addressee Estimation to foster robots' socio-perceptual skills based on the Shared Perception framework

As defined in Chapter 2, Shared Perception moves from the perception of others to achieve an augmented perception of the environment. Others' intentional reference toward the environment is perceived, interpreted, and employed as an additional source of information to enhance one's perception. The body of the other is a crucial element for Shared Perception to emerge. Its intentional reference toward the environment is grasped from its body: attention, perspective, actions, and inner states are all achievable by perceiving and interpreting bodily non-verbal behaviors.

Addressee Estimation has been developed as a Shared Perception skill starting from here. As the ability to perceive whom a person is talking to, it is inherently connected to the concept of intention understanding. Rarely, the addressee of a speaker is defined as overt information, explicated in the speech. More often, it is made clear by the context or the non-verbal behavior of the speaker: elements which bring to interpret the implicit goal of the person talking. Endowing the robot with the ability to interpret the non-verbal behavior of the speaker and localize its addressee serves as a basis for Shared Perception to occur and to foster social interaction. Beyond enabling the robot to engage in conversation with others and evaluate if being approached, it allows understanding the social dynamics of the group, contextualizing the content of a speech, discovering not-yet-detected agents in the environment, etc.

With respect to previous literature on the topic, the Addressee Position Estimation model developed in this thesis moves from a binary prediction approach to understand if the robot was addressed, toward predicting the position of the addressee with respect to the speaker. From this perspective, the model yields information about the position of the addressee starting from the bodily behavior of the speaker. Specifically, the two features that disclose the speaker's intention are the head direction and the body pose. The former served as a proxy for the speaker's attention, the latter provided information about the speaker's perspective (body direction) and, to a lesser and not directly explored extent, actions (body movements). The results of the model suggested this to be sufficient information to predict the addressee's position. No other contextual knowledge about the number or position of third agents in the room was needed to achieve sufficiently reliable results, as it was, instead, in several previous models of Addressee Estimation.

Shared Perception is based on the perception of others' intentional relation toward the environment but is fulfilled with the integration of two other sources of information: the self and the environment. However fundamental to eliciting Shared Perception, the APE model does not reach a complete Shared Perception. For this to happen, the whole Addressee Estimation architecture designed at the end of Chapter 5 is needed. This way it will be possible to move from the perception of the speaker's body to the other benefits enabled by Shared Perception: understanding others, creating a common ground, and augmenting perception of the external environment, as described in Section 5.5. Future work is therefore needed to finalize the model of Addressee Estimation as Shared Perception. Though the architecture has already been designed in most components, the Addressee Estimation modules still need to be implemented on the iCub robot and connected to those preexistent.

6.2 Final Overview

The present work emerges at the intersection of human-robot interaction, cognitive robotics, and social robotics. From these disciplines, the thesis respectively inherits the quest for a natural and efficient interaction between humans and robots, the aim to foster robots' autonomy with cognitive abilities, and the social dimension of the interaction. Accordingly, the motivation for this research has been the need to improve the interactive social abilities of robots to make them effective instruments and collaborators of human development and well-being. The theoretical and experimental investigation of human perception as well as the modeling and development of a perceptual skill inspired by it were driven by this need.

Shared Perception offers a plausible interpretative account of human perception during social interaction and, for this reason, provides support to improve socio-cognitive HRI:

- it is a reference for designing human-scaled interactions
- it promotes the implementation of human-inspired cognitive and perceptual skills.

If robots are developed to help and assist humans, designing interactions with them based on human social abilities appears straightforward. Principles of human perception have already been used to develop technological devices. For instance, the perceptual laws identified by Gestalt Theory (see Section 2.3.1) have been adapted for web page [252] and, more generally, interactive media design [86], with the objective to enhance the user experience (UX) by modeling the device on human perception. The role of Shared Perception in HRI could follow the same direction. Based on criteria inspired by human social perception, the design of the robot's body, movements, and gestures, as well as the structure of proxemic and context, could be all elements to improve the quality of the interaction in terms of smoothness, engagement, and effectiveness. The results of the experiment described in Chapter 3 promote this outlook. In that case, we designed the interaction using a humanoid robot expressing biological motion to obtain a higher attribution of anthropomorphism, the robot's head movements, oculomotor behavior, and emotion-expression with eyelids to elicit attribution of intentionality, mutual gaze and smiles to evoke pro-sociality, speech to foster engagement. As a result, despite the identical visual stimulus, the interaction with the social or mechanical robot induced an alteration in visual perception of space, leading participants to rely on the shared visual stimuli more than on their private internal models while interacting with the social robot.

In embodied artificial systems, human awareness can be considered the system's capability to interpret human behavior and be sensitive to their inner states, such as beliefs, emotions, or intentions. Shared Perception indicates that this ability is crucial at several levels: it increases the understanding of others, aids the robot in discovering the environment, and founds interaction on common ground. For these reasons, this thesis suggests that the dynamics of Shared Perception, based on the other, the self, and the environment, should be considered when it comes to developing perceptual and cognitive skills for autonomous robots. Addressee Estimation was one possible solution to practically implement such dynamics. I decided to tackle the Addressee Estimation problem because it is extremely helpful in multi-party interaction and can support the robot's capability to infer the presence of new people in the environment. Two skills that would improve robots' awareness and perception of social environments. However, the same logic underlying the model development reported

in Chapter 5 can be applied to a number of other situations. Every time the behavior of others can reveal, for instance, hidden areas of the environment, subjective qualities of objects, the presence of new people, or specific relations among people, Shared Perception can be of help. Improving the robot's awareness and perception of the environment thanks to other people would not only make it more autonomous but may profoundly improve the interaction.

These three years of research have been guided by the idea that to interact, hence to act together, two partners need to perceive together. On the same line, to perceive together each partner should integrate personal information, information disclosed by the other, and information from the environment: in two words, *share perception*.

Publications

Listed below are the publications related to the contribution of the thesis or made during my Ph.D. period:

Book chapter

- **Mazzola C.***, Incao S.*, Rea F., Sciutti A., Marassi M., "Human Experience And Robotic Experience: A Reciprocal Exchange Of Perspectives", in *Humane Robotics. A Multidisciplinary Approach Towards the Development of Humane-Centered Technologies*, Riva G., Marchetti A. (Eds.), Publisher: Vita e Pensiero, 2022

Journal Articles

- **Mazzola C.**, Rea F. and Sciutti A., "Shared perception is different from individual perception: a new look on context dependency," in IEEE Transactions on Cognitive and Developmental Systems, 2022, doi: 10.1109/TCDS.2022.3185100.

- Tsfasman M.*, Philippsen A.*, **Mazzola C.**, Thill S., Sciutti A., Nagai Y., "The world seems different in a social context: A neural network analysis of human experimental data", in PLOS ONE, 17(8), 2022, doi: 10.1371/journal.pone.0273643

- Incao S.*, **Mazzola C.**, Sciutti A., "The impact of early aging on visual perception of space and time", in Frontiers in Human Neuroscience, 2022, doi: 10.3389/fnhum.2022.988644.

International conferences

- **Mazzola C.***, Aroyo A.M., Rea F., Sciutti A. "Interacting with a Social Robot Affects Visual Perception of Space", in Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), Association for Computing Machinery, New York, NY, USA, 549–557, 2020,doi: 10.1145/3319502.3374819.

- **Mazzola C.***, Incao S., Marassi M., Rea F., Sciutti A., "A hermeneutical approach to provide robots with socially adaptive perception", in Social Robots in Social Institutions. Proceedings of Robophilosophy 2022, Hakli, R., Mäkelä, P., Seibt, J. (Eds.), IOS Press, Helsinki, 2022, doi: 10.3233/FAIA220634.

- Tonelli A.*, **Mazzola C.**, Sciutti A., Gori M., "The influence of visual experience on context dependency in hearing distance estimation", International Multisensory Forum Research, 2022.

Workshops

- Bavazzano I.,**Mazzola C.**, Belgiovine G., Casadio M., Sciutti A., "Context dependency in a social rehabilitation scenario", IEEE ICDL 2021 Workshop Spatio-temporal Aspects of Embodied Predictive Processing. (ICDL Workshop StEPP'21), 2021, doi: 10.5281/zenodo.5579180

- Venturini F.*, **Mazzola C.**, Marassi M., "A specific role for Damasio's Somatic Markers in artificial decision-making: advantages and potentials for future implementations", IEEE ICDL 2021 Workshop Feel-COG – the role of affect in the development of cognition, 2021.

References

[1] Adams, R. A., Shipp, S., and Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function*, 218(3):611–643.

[2] Addison, A., Bartneck, C., and Yogeeswaran, K. (2019). Robots can be more than black and white: Examining racial bias towards robots. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 493–498, New York, NY, USA. Association for Computing Machinery.

[3] Admoni, H. and Scassellati, B. (2017). Social eye gaze in human-robot interaction: A review. *J. Hum.-Robot Interact.*, 6(1):25–63.

[4] Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. (2012). Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In Esposito, A., Esposito, A. M., Vinciarelli, A., Hoffmann, R., and Müller, V. C., editors, *Cognitive Behavioural Systems*, pages 114–130, Berlin, Heidelberg. Springer Berlin Heidelberg.

[5] Anzalone, S. M., Boucenna, S., Ivaldi, S., and Chetouani, M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics*, 7(4):465–478.

[6] Anzalone, S. M., Xavier, J., Boucenna, S., Billeci, L., Narzisi, A., Muratori, F., Cohen, D., and Chetouani, M. (2019). Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment. *Pattern Recognition Letters*, 118:42–50.

[7] Apperly, I. A. and Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116:953–970.

[8] Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, 63(4), 596.

[9] Auer, P. (2018). *Gaze, addressee selection and turn-taking in three-party interaction*, pages 197–231. John Benjamins Amsterdam.

[10] Bach, P. and Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass*, 11(7):e12312.

[11] Bakx, I., Van Turnhout, K., and Terken, J. (2003). Facial orientation during multi-party interaction with information kiosks. *Proceedings of INTERACT 2003 ZÃ1/4rich, Switzerland*, pages 163–170.

[12] Banks, J. (2020). Theory of mind in social robots: replication of five established human tests. *International Journal of Social Robotics*, 12(2):403–414.

[13] Barnes-Holmes, Y., McHugh, L., and Barnes-Holmes, D. (2004). Perspective-taking and theory of mind: A relational frame account. *The Behavior Analyst Today*, 5(1):15.

[14] Baron-Cohen, S. (1995). *The eye direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology*, pages pp. 41–59. Inc. Lawrence Erlbaum Associates.

[15] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). The autism-spectrum quotient (aq): Evidence from asperger syndrome / high-functioning autism , males and females , scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31.

[16] Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., and Sabanović, S. (2020). *Human-robot interaction: An introduction*. Cambridge University Press.

[17] Bartneck, C., Kuli, D., and Croft, E. (2009). Measurement instruments for the anthropomorphism , animacy , likeability , perceived intelligence , and perceived safety of robots. *International Journal of Social Robotics*, 1:71–81.

[18] Bejjanki, V. R., Knill, D. C., and Aslin, R. N. (2016). Learning and inference using complex generative models in a spatial localization task. *Journal of Vision*, 16(5):9–9.

[19] Belgiovine, G., Gonzalez-Billandon, J., Sandini, G., Rea, F., and Sciutti, A. (2022). Towards an hri tutoring framework for long-term personalization and real-time adaptation. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22 Adjunct, page 139–145, New York, NY, USA. Association for Computing Machinery.

[20] Belgiovine, G., Rea, F., Barros, P., Zenzeri, J., and Sciutti, A. (2020). Sensing the partner: Toward effective robot tutoring in motor skill learning. In Wagner, A. R., Feil-Seifer, D., Haring, K. S., Rossi, S., Williams, T., He, H., and Sam Ge, S., editors, *Social Robotics*, pages 296–307, Cham. Springer International Publishing.

[21] Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*, 3(21):eaat5954.

[22] Bertenthal, B. (1996). Origins and early development of perception, action, and representation. *Annual review of psychology*, 47(1):431–459.

[23] Bicks, L. K., Koike, H., Akbarian, S., and Morishita, H. (2015). Prefrontal cortex and social cognition in mouse and man. *Frontiers in psychology*, 6:1805.

[24] Bolis, D. and Schilbach, L. (2020). 'I interact therefore I am': The self as a historical product of dialectical attunement. *Topoi*, 39(3):521–534.

[25] Brooks, R. and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, 8(6):535–543.

[26] Brown, E. and Brüne, M. (2012). The role of prediction in social neuroscience. *Frontiers in Human Neuroscience*, 6:147.

[27] Brown, L. N. and Howard, A. M. (2014). The positive effects of verbal encouragement in mathematics education using a social robot. In *2014 IEEE Integrated STEM Education Conference*, pages 1–5.

[28] Bubic, A., von Cramon, D. Y., and Schubotz, R. I. (2010). Prediction, cognition and the brain. *Front Hum Neurosci*, 4:25–25. 20631856[pmid].

[29] Bushnell, E. W. and Boudreau, J. P. (1993). Motor development and the mind: The potential role of motor abilities as a determinant of aspects of perceptual development. *Child Development*, 64(4):1005–1021.

[30] Butterworth, G. (1991). *The ontogeny and phylogeny of joint visual attention*. Basil Blackwell.

[31] Cangelosi, A. and Asada, M. (2022). *Cognitive robotics*. MIT Press.

[32] Cannon-Bowers, J. A. and Salas, E. (2001). Reflections on shared cognition. *Journal of Organizational Behavior*, 22(2):195–202.

[33] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[34] Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., and Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, pages i–174.

[35] Chevalier, P., Kompatsiari, K., Ciardo, F., and Wykowska, A. (2020). Examining joint attention with the use of humanoid robots-a new approach to study fundamental mechanisms of social cognition. *Psychonomic Bulletin & Review*, 27(2):217–236.

[36] Chiorri, C., Bracco, F., Piccinno, T., Modafferi, C., and Battini, V. (2014). Psychometric properties of a revised version of the ten item personality inventory. *European Journal of Psychological Assessment*.

[37] Chita-Tegmark, M., Lohani, M., and Scheutz, M. (2019). Gender effects in perceptions of robots and humans with varying emotional intelligence. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 230–238.

[38] Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., and Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. *Journal of Neuroscience*, 32:1056–60.

[39] Cifuentes, C. A., Pinto, M. J., Céspedes, N., and Múnera, M. (2020). Social robots in therapy and care. *Current Robotics Reports*, 1(3):59–74.

[40] Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36:181–204.

[41] Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

[42] Conway, J. R., Lee, D., Ojaghi, M., Catmur, C., and Bird, G. (2017). Submentalizing or mentalizing in a level 1 perspective-taking task: A cloak and goggles test. *Journal of Experimental Psychology: Human Perception and Performance*, 43(3):454.

[43] Corlett, P. R., Frith, C. D., and Fletcher, P. C. (2009). From drugs to deprivation: a bayesian framework for understanding models of psychosis. *Psychopharmacology*, 206(4):515–530.

[44] Creem-Regehr, S. H., Gagnon, K. T., Geuss, M. N., and Stefanucci, J. K. (2013). Relating spatial perspective taking to the perception of other's affordances: Providing a foundation for predicting the future behavior of others. *Frontiers in Human Neuroscience*, 7:596.

[45] Crockett, M. J. and Fehr, E. (2014). Social brains on drugs: tools for neuromodulation in social neuroscience. *Social cognitive and affective neuroscience*, 9(2):250–254.

[46] Davidson, D. (1991). Three varieties of knowledge. *Royal Institute of Philosophy Supplement*, 30:153–166.

[47] De Jaegher, H., Di Paolo, E., and Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences*, 14(10):441–447.

[48] Den Ouden, H., Kok, P., and De Lange, F. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, 3:548.

[49] Di Cesare, G., Di Dio, C., Marchi, M., and Rizzolatti, G. (2015). Expressing our internal states and understanding those of others. *Proceedings of the National Academy of Sciences*, 112(33):10331–10335.

[50] Di Cesare, G., Di Dio, C., Rochat, M., Sinigaglia, C., Bruschweiler-Stern, N., Stern, D., and Rizzolatti, G. (2014). The neural correlates of 'vitality form'recognition: an fmri study: This work is dedicated to daniel stern, whose immeasurable contribution to science has inspired our research. *Social cognitive and affective neuroscience*, 9(7):951–960.

[51] Di Cesare, G., Vannucci, F., Rea, F., Sciutti, A., and Sandini, G. (2020). How attitudes generated by humanoid robots shape human brain activity. *Scientific Reports*, 10(1):1–12.

[52] Divekar, R. R., Kephart, J. O., Mou, X., Chen, L., and Su, H. (2019). You talkin' to me? a practical attention-aware embodied agent. In Lamas, D., Loizides, F., Nacke, L., Petrie, H., Winckler, M., and Zaphiris, P., editors, *Human-Computer Interaction – INTERACT 2019*, pages 760–780, Cham. Springer International Publishing.

[53] Doğan, F. I., Gillet, S., Carter, E. J., and Leite, I. (2020). The impact of adding perspective-taking to spatial referencing during human–robot interaction. *Robotics and Autonomous Systems*, 134:103654.

[54] Duch, W., Oentaryo, R. J., and Pasquier, M. (2008). Cognitive architectures: Where do we go from here? In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, volume 171, pages 122–136.

[55] Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., and Garnero, L. (2010). Inter-brain synchronization during social interaction. *PloS one*, 5(8):e12166.

[56] Echterhoff, G., Higgins, E. T., and Levine, J. M. (2009). Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science*, 4(5):496–521.

[57] Echterhoff, G. and Kopietz, R. (2018). The socially shared nature of memory: From joint encoding to communication. *Collaborative Remembering: Theories, Research, and Applications*, pages 113–134.

[58] Falck-Ytter, T., Gredebäck, G., and Hofsten, C. V. (2006). Infants predict other people's action goals. *Nature Neuroscience*, 9:878–879.

[59] Feil-Seifer, D. and Mataric, M. (2005). Defining socially assistive robotics. In *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, pages 465–468.

[60] Ferrari, F., Paladino, M., and Jetten, (2016). Blurring human – machine distinctions : Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, 8:287–302.

[61] Ferrari, P. F. and Coudé, G. (2018). Mirror neurons, embodied emotions, and empathy. In *Neuronal correlates of empathy*, pages 67–77. Elsevier.

[62] Fiebich, A. and Gallagher, S. (2013). Joint attention in joint action. *Philosophical Psychology*, 26(4):571–587.

[63] Flavell, J. H. (1977). The development of knowledge about visual perception. In *Nebraska symposium on motivation*. University of Nebraska Press.

[64] Flavell, J. H., Everett, B. A., Croft, K., and Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the level 1-level 2 distinction. *Developmental Psychology*, 17:99–103.

[65] Frampton, M., Fernández, R., Ehlen, P., Christoudias, C. M., Darrell, T., and Peters, S. (2009). Who is "you"? combining linguistic and gaze features to resolve second-person references in dialogue. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 273–281.

[66] Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1211–1221.

[67] Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLOS ONE*, 4:1–13.

[68] Furlanetto, T., Becchio, C., Samson, D., and Apperly, I. (2016). Altercentric interference in level 1 visual perspective taking reflects the ascription of mental states, not submentalizing. *Journal of Experimental Psychology: Human Perception and Performance*, 42:158–163.

[69] Gadamer, H.-G. (2013). *Truth and method*. A&C Black.

[70] Gaggioli, A., Chirico, A., Di Lernia, D., Maggioni, M. A., Malighetti, C., Manzi, F., Marchetti, A., Massaro, D., Rea, F., Rossignoli, D., Sandini, G., Villani, D., Wiederhold, B. K., Riva, G., and Sciutti, A. (2021). Machines like us and people like you: Toward human–robot shared experience. *Cyberpsychology, Behavior, and Social Networking*, 24(5):357–361.

[71] Gallagher, S. (1986). Body image and body schema: A conceptual clarification. *The Journal of mind and behavior*, pages 541–554.

[72] Gao, Z., Shi, Q., Fukuda, T., Li, C., and Huang, Q. (2019). An overview of biomimetic robots with animal behaviors. *Neurocomputing*, 332:339–350.

[73] Gardner, M. R., Hull, Z., Taylor, D., and Edmonds, C. J. (2018). 'Spontaneous' visual perspective-taking mediated by attention orienting that is voluntary and not reflexive. *Quarterly journal of experimental psychology (2006)*, 71:1020–1029.

[74] Garello, L., Lastrico, L., Rea, F., Mastrogiovanni, F., Noceti, N., and Sciutti, A. (2021). Property-aware robot object manipulation: a generative approach. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–7.

[75] Gargot, T., Asselborn, T., Zammouri, I., Brunelle, J., Johal, W., Dillenbourg, P., Archambault, D., Chetouani, M., Cohen, D., and Anzalone, S. M. (2021). "it is not the robot who learns, it is me." treating severe dysgraphia using child–robot interaction. *Frontiers in Psychiatry*, 12.

[76] Gazzola, V., Rizzolatti, G., Wicker, B., and Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage*, 35(4):1674–1684.

[77] Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419.

[78] Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition (1st ed.)*. Psychology press.

[79] Gillet, S. and Leite, I. (2020). A robot mediated music mixing activity for promoting collaboration among children. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 212–214, New York, NY, USA. Association for Computing Machinery.

[80] Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement_3):15647–15654.

[81] Goffman, E. (1981). *Forms of talk*. University of Pennsylvania Press.

[82] Gonzalez-Billandon, J., Belgiovine, G., Sciutti, A., Sandini, G., and Francesco, R. (2021a). Towards a cognitive framework for multimodal person recognition in multiparty hri. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, page 412–416, New York, NY, USA. Association for Computing Machinery.

[83] Gonzalez-Billandon, J., Belgiovine, G., Tata, M., Sciutti, A., Sandini, G., and Rea, F. (2021b). Self-supervised learning framework for speaker localisation with a humanoid robot. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–7.

[84] Gonzalez-Billandon, J., Sciutti, A., Tata, M., Sandini, G., and Rea, F. (2020). Audiovisual cognitive architecture for autonomous learning of face localisation by a humanoid robot. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5979–5985.

[85] Gori, M., Sciutti, A., Burr, D., and Sandini, G. (2011). Direct and indirect haptic calibration of visual size judgments. *PLoS One*, 6(10):e25599.

[86] Graham, L. (2008). Gestalt theory in interactive media design. *Journal of Humanities & Social Sciences*, 2(1).

[87] Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315:619–619.

[88] Gzesh, S. M. and Surber, C. F. (1985). Visual perspective-taking skills in children. *Child Development*, 56:1204–1213.

[89] Haddadin, S. (2020). *Physical Human-Robot Interaction*. Springer Berlin Heidelberg.

[90] Hari, R., Henriksson, L., Malinen, S., and Parkkonen, L. (2015). Centrality of social interaction in human brain function. *Neuron*, 88(1):181–193.

[91] Hari, R. and Kujala, M. V. (2009). Brain basis of human social interaction: From concepts to brain imaging. *Physiological Reviews*, 89(2):453–479.

[92] Helmholtz, H. (1866). *Handbuch der physiologischen Optik*.

[93] Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7:461–469.

[94] Horiguchi, S., Kanda, N., and Nagamatsu, K. (2019). Multimodal response obligation detection with unsupervised online domain adaptation. In *INTERSPEECH*, pages 4180–4184.

[95] Huang, H.-H., Baba, N., and Nakano, Y. (2011). Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation using nonverbal information. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, page 401–408, New York, NY, USA. Association for Computing Machinery.

[96] Hudson, M., Nicholson, T., Ellis, R., and Bach, P. (2016). I see what you say: Prior knowledge of other's goals automatically biases the perception of their actions. *Cognition*, 146:245–250.

[97] Husserl, E. (1975). *Logische Untersuchungen*. Holenstein Elmar, Nijhoff, Den Haag.

[98] Huttenlocher, J., Hedges, L. V., and Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of experimental psychology: General*, 129(2):220.

[99] Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., and Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS biology*, 3(3):e79.

[100] Ingle, D., Marcus, N., and Johal, W. (2021). The valley of non-distraction: Effect of robot's human-likeness on perception load. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, page 99–103, New York, NY, USA. Association for Computing Machinery.

[101] Insel, T. R. and Fernald, R. D. (2004). How the brain processes social information: searching for the social brain. *Annu. Rev. Neurosci.*, 27:697–722.

[102] Ishiguro, H. (2020). *Android Science and Engineering*. Springer Berlin Heidelberg.

[103] Ishii, R., Otsuka, K., Kumano, S., and Yamato, J. (2016). Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. 6(1).

[104] Jakobson, R. (1981). *Linguistics and Poetics*, pages 18–51. De Gruyter Mouton, Berlin, Boston.

[105] Jakobson, R. (1987). *Language in literature*. Harvard University Press.

[106] jamovi Project, T. (2020). jamovi (version 1.6.1) [computer software].

[107] Jayagopi, D. B., Sheikhi, S., Klotz, D., Wienke, J., Odobez, J.-M., Wrede, S., Khalidov, V., Nguyen, L., Wrede, B., and Gatica-Perez, D. (2012). The vernissage corpus: A multimodal human-robot-interaction dataset. page 8.

[108] Jayagopi, D. B., Sheiki, S., Klotz, D., Wienke, J., Odobez, J.-M., Wrede, S., Khalidov, V., Nyugen, L., Wrede, B., and Gatica-Perez, D. (2013). The vernissage corpus: A conversational human-robot-interaction dataset. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 149–150.

[109] Jazayeri, M. and Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature neuroscience*, 13:1020–6.

[110] Johansson, M. and Skantze, G. (2015). Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 305–314.

[111] Jovanovic, N., op den Akker, R., and Nijholt, A. (2006). Addressee identification in face-to-face meetings. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 169–176.

[112] Kaiser, M. K. and Proffitt, D. R. (1984). The development of sensitivity to causally relevant dynamic information. *Child Development*, pages 1614–1624.

[113] Karaminis, T., Cicchini, G. M., Neil, L., Cappagli, G., Aagten-Murphy, D., Burr, D., and Pellicano, E. (2016). Central tendency effects in time interval reproduction in autism. *Scientific Reports*, 6:1–13.

[114] Katzenmaier, M., Stiefelhagen, R., and Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, page 144–151, New York, NY, USA. Association for Computing Machinery.

[115] Kelley, M. S., Noah, J. A., Zhang, X., Scassellati, B., and Hirsch, J. (2021). Comparison of human social brain activity during eye-contact with another human and a humanoid robot. *Frontiers in Robotics and AI*, 7:599581.

[116] Kennedy, J., Baxter, P., and Belpaeme, T. (2015). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 67–74. IEEE.

[117] Kersten, D. and Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, 13(2):150–158.

[118] Kim, M., Kwon, T., and Kim, K. (2018). Can human–robot interaction promote the same depth of social information processing as human–human interaction? *International Journal of Social Robotics*, 10:33–42.

[119] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

[120] Knoblich, G. and Sebanz, N. (2006). The social nature of perception and action. *Current directions in psychological science*, 15(3):99–104.

[121] Köhler, W. (1929). *Gestalt Psychology*. New York: Liveright.

[122] Kompatsiari, K., Ciardo, F., Tommaso, D. D., and Wykowska, A. (2019). Measuring engagement elicited by eye contact in human-robot interaction. *IEEE International Conference on Intelligent Robots and Systems*, pages 6979–6985.

[123] Kompatsiari, K., Perez-Osorio, J., Tommaso, D. D., Metta, G., and Wykowska, A. (2018). Neuroscientifically-grounded research for improved human-robot interaction. *IEEE International Conference on Intelligent Robots and Systems*, pages 3403–3408.

[124] Kotseruba, I. and Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1):17–94.

[125] Kurt, K. (1935). *Principles of Gestalt Psychology*. New York: Harcourt, Brace.

[126] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82:1–26.

[127] Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64.

[128] Lakatos, G., Wood, L. J., Syrdal, D. S., Robins, B., Zaraki, A., and Dautenhahn, K. (2021). Robot-mediated intervention can assist children with autism to develop visual perspective taking skills. *Paladyn, Journal of Behavioral Robotics*, 12(1):87–101.

[129] Lamm, C., Bukowski, H., and Silani, G. (2016). From shared to distinct self–other representations in empathy: evidence from neurotypical function and socio-cognitive disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371.

[130] Lanillos, P., Oliva, D., Philippsen, A., Yamashita, Y., Nagai, Y., and Cheng, G. (2020). A review on neural network models of schizophrenia and autism spectrum disorder. *Neural Networks*, 122:338–363.

[131] Lastrico, L., Garello, L., Rea, F., Noceti, N., Mastrogiovanni, F., Sciutti, A., and Carfì, A. (2022). Robots with different embodiments can express and influence carefulness in object manipulation. In *2022 IEEE International Conference on Development and Learning (ICDL)*, pages 280–286.

[132] Lee, J. J., Sha, F., and Breazeal, C. (2019). A bayesian theory of mind approach to nonverbal communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 487–496.

[133] Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37.

[134] Liepelt, R., Cramon, D., and Brass, M. (2008). What is matched in direct matching? intention attribution modulates motor priming. *Journal of Experimental Psychology: human perception and performance*, 34(3):578.

[135] Lüdecke, D. (2021). sjstats: Statistical functions for regression models (version 0.18.1).

[136] Malik, U., Barange, M., Saunier, J., and Pauchet, A. (2019). Using multimodal information to enhance addressee detection in multiparty interaction. In *ICAART (1)*, pages 267–274.

[137] Malik, U., Barange, M., Saunier, J., and Pauchet, A. (2021). A novel focus encoding scheme for addressee detection in multiparty interaction using machine learning algorithms. *Journal on Multimodal User Interfaces*, 15(2):175–188.

[138] Martin, T. and Schwartz, D. L. (2005). Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts. *Cognitive Science*, 29:587–625.

[139] Matarese, M., Rea, F., and Sciutti, A. (2022). Perception is only real when shared: A mathematical model for collaborative shared perception in human-robot interaction. *Frontiers in Robotics and AI*, 9.

[140] Maturana, H. R. and Varela, F. J. (1987). *The tree of knowledge: The biological roots of human understanding.* New Science Library/Shambhala Publications.

[141] Mazzola, C., Aroyo, A. M., Rea, F., and Sciutti, A. (2020). Interacting with a social robot affects visual perception of space. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 549–557, New York, NY, USA. Association for Computing Machinery.

[142] Mazzola, C., Rea, F., and Sciutti, A. (2022). Shared perception is different from individual perception: a new look on context dependency. *IEEE Transactions on Cognitive and Developmental Systems*.

[143] McMillan, D., Brown, B., Kawaguchi, I., Jaber, R., Solsona Belenguer, J., and Kuzuoka, H. (2019). Designing with gaze: Tama – a gaze activated smart-speaker. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

[144] Meltzoff, A. N. (2007). 'like me': a foundation for social cognition. *Developmental Science*, 10(1):126–134.

[145] Meltzoff, A. N., Brooks, R., Shon, A. P., and Rao, R. P. (2010). "social" robots are psychological agents for infants: A test of gaze following. *Neural networks*, 23(8-9):966–972.

[146] Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Librairie Gallimard.

[147] Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., Bernardino, A., and Montesano, L. (2010). The icub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, 23(8):1125–1134.

[148] Minh, T. L., Shimizu, N., Miyazaki, T., and Shinoda, K. (2018). Deep learning based multi-modal addressee recognition in visual scenes with utterances. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1546–1553. International Joint Conferences on Artificial Intelligence Organization.

[149] Moll, H. and Meltzoff, A. N. (2011). Joint attention as the fundamental basis of understanding perspectives. In Seeman, A., editor, *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience*, chapter 15, pages 393–413. MIT Press, Oxford.

[150] Morgado, N., Muller, D., Gentaz, E., and Palluel-Germain, R. (2011). Close to me? the influence of affective closeness on space perception. *Perception*, 40:877–879.

[151] Morgado, N., Muller, D., Pinelli, M., Éric Guinet, Édouard Gentaz, and Palluel-Germain, R. (2013). Do friendship influence space perception? with particular reference to the curse of the suspicious participants. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35:1–6.

[152] Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100.

[153] Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., and Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child Development*, 78(3):938–954.

[154] Mundy, P. and Newell, L. (2007). Attention, joint attention, and social cognition. *Current Directions in Psychological Science*, 16(5):269–274.

[155] Murata, S., Namikawa, J., Arie, H., Sugano, S., and Tani, J. (2013). Learning to reproduce fluctuating time series by inferring their time-dependent stochastic properties: Application in robot learning via tutoring. *IEEE Transactions on Cognitive and Developmental Systems*, 5(4):298–310.

[156] Möller, R., Furnari, A., Battiato, S., Härmä, A., and Farinella, G. M. (2021). A survey on human-aware robot navigation. *Robotics and Autonomous Systems*, 145:103837.

[157] Nagai, Y. (2019). Predictive learning: its key role in early cognitive development. *Philosophical Transactions of the Royal Society B*, 374(1771):20180030.

[158] Nagai, Y. (2020). *Predictive Coding for Cognitive Development*, pages 1–8. Springer Berlin Heidelberg, Berlin, Heidelberg.

[159] Nakisa, B., Rastgoo, M. N., Rakotonirainy, A., Maire, F., and Chandran, V. (2020). Automatic emotion recognition using temporal multimodal deep learning. *IEEE Access*, 8:225463–225474.

[160] Oberman, L. M., McCleery, J. P., Ramachandran, V. S., and Pineda, J. A. (2007). Eeg evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*, 70(13-15):2194–2203.

[161] Oliva, D., Philippsen, A., and Nagai, Y. (2019). How development in the Bayesian brain facilitates learning. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE.

[162] op den Akker, H. and op den Akker, R. (2009). Are you being addressed?-real-time addressee detection to support remote participants in hybrid meetings. In *Proceedings of the SIGDIAL 2009 Conference*, pages 21–28.

[163] O'Grady, C., Scott-Phillips, T., Lavelle, S., and Smith, K. (2020). Perspective-taking is spontaneous but not automatic. *Quarterly Journal of Experimental Psychology*, 73(10):1605–1628.

[164] Paetzel, M., Perugia, G., and Castellano, G. (2020). The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 73–82, New York, NY, USA. Association for Computing Machinery.

[165] Palinko, O., Rea, F., Sandini, G., and Sciutti, A. (2016). Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054.

[166] Park, H. W., Grover, I., Spaulding, S., Gomez, L., and Breazeal, C. (2019). A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):687–694.

[167] Pellicano, E. and Burr, D. (2012). When the world becomes ' too real ': a bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16.

[168] Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., and Pioggia, G. (2016). Autism and social robotics: A systematic review. *Autism Research*, 9(2):165–183.

[169] Petzschner, F. H., Glasauer, S., and Stephan, K. E. (2015). A bayesian perspective on magnitude estimation. *Trends in cognitive sciences*, 19(5):285–293.

[170] Pezzulo, G., Barsalou, L., Cangelosi, A., Fischer, M., McRae, K., and Spivey, M. (2013). Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in Psychology*, 3.

[171] Pezzulo, G., Parr, T., and Friston, K. (2022). The evolution of brain architectures for predictive coding and active inference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1844):20200531.

[172] Philippsen, A. and Nagai, Y. (2020a). Deficits in prediction ability trigger asymmetries in behavior and internal representation. *Frontiers in Psychiatry*, 11:1–16.

[173] Philippsen, A. and Nagai, Y. (2020b). A predictive coding account for cognition in human children and chimpanzees: A case study of drawing. *IEEE Transactions on Cognitive and Developmental Systems*.

[174] Piaget, J. (1956). *The Child's Conception of Space*. Routledge.

[175] Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

[176] Ramachandram, D. and Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

[177] Recasens, A., Khosla, A., Vondrick, C., and Torralba, A. (2015). Where are they looking? In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

[178] Richardson, D. C., Street, C. N., Tan, J. Y., Kirkham, N. Z., Hoover, M. A., and Cavanaugh, A. G. (2012). Joint perception: Gaze and social context. *Frontiers in Human Neuroscience*, 6:1–8.

[179] Richir, M. (2000). *Phénoménologie en esquisses : nouvelles fondations*. J. Millon.

[180] Richter, V., Carlmeyer, B., Lier, F., Meyer zu Borgsen, S., Schlangen, D., Kummert, F., Wachsmuth, S., and Wrede, B. (2016). Are you talking to me? improving the robustness of dialogue systems in a multi party hri scenario by incorporating gaze direction and lip movement of attendees. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, HAI '16, page 43–50, New York, NY, USA. Association for Computing Machinery.

[181] Ritter, F. E., Tehranchi, F., and Oury, J. D. (2019). Act-r: A cognitive architecture for modeling cognition. *WIREs Cognitive Science*, 10(3).

[182] Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192.

[183] Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141.

[184] Roach, N. W., McGraw, P. V., Whitaker, D. J., and Heron, J. (2017). Generalization of prior information for rapid bayesian time estimation. *Proceedings of the National Academy of Sciences*, 114(2):412–417.

[185] Romeo, M., Hernández García, D., Jones, R., and Cangelosi, A. (2019). Deploying a deep learning agent for hri with potential "end-users" at multiple sheltered housing sites. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, HAI '19, page 81–88, New York, NY, USA. Association for Computing Machinery.

[186] Romeo, M., Hernández García, D., Han, T., Cangelosi, A., and Jokinen, K. (2021). Predicting apparent personality from body language: benchmarking deep learning architectures for adaptive social human–robot interaction. *Advanced Robotics*, 35(19):1167–1179.

[187] Roncone, A., Pattacini, U., Metta, G., and Natale, L. (2016). A cartesian 6-dof gaze controller for humanoid robots. In *Robotics: science and systems*, volume 2016.

[188] Rubin, E. (1915). *Visual perception of figures: studies in psychological analysis*. Copenhagen: Gyldendahl.

[189] Ruta, L., Mazzone, D., and Mazzone, L. (2012). The autism-spectrum quotient — italian version : A cross- cultural confirmation of the broader autism phenotype. *Journal of autism and developmental disorders*, 42:625–633.

[190] Salam, H. and Chetouani, M. (2015). Engagement detection based on mutli-party cues for human robot interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 341–347.

[191] Salatas, H. and Flavell, J. H. (1976). Development of two components of knowledge. *Child Development*, 47:103–109.

[192] Sandini, G. and Sciutti, A. (2018). Humane robots—from robots with a humanoid body to robots with an anthropomorphic mind. *J. Hum.-Robot Interact.*, 7(1).

[193] Sandini, G., Sciutti, A., and Vernon, D. (2020). *Cognitive Robotics*. Springer Berlin Heidelberg.

[194] Scassellati, B., Boccanfuso, L., Huang, C.-M., Mademtzi, M., Qin, M., Salomons, N., Ventola, P., and Shic, F. (2018). Improving social skills in children with asd using a long-term, in-home social robot. *Science Robotics*, 3(21):eaat7544.

[195] Sciutti, A., Burr, D., Saracco, A., Sandini, G., and Gori, M. (2014a). Development of context dependency in human space perception. *Experimental Brain Research*, 232:3965–3976.

[196] Sciutti, A., Mara, M., Tagliasco, V., and Sandini, G. (2018). Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine*, 37(1):22–29.

[197] Sciutti, A., Patanè, L., Nori, F., and Sandini, G. (2014b). Understanding object weight from human and humanoid lifting actions. *IEEE Transactions on Autonomous Mental Development*, 6:80–92.

[198] Sciutti, A. and Sandini, G. (2017). Interacting with robots to investigate the bases of social interaction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(12):2295–2304.

[199] Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in cognitive sciences*, 10(2):70–76.

[200] Sheikhi, S., Babu Jayagopi, D., Khalidov, V., and Odobez, J.-M. (2013). Context aware addressee estimation for human robot interaction. In *GazeIn '13: Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*, GazeIn '13, page 1–6, New York, NY, USA. Association for Computing Machinery.

[201] Sheikhi, S. and Odobez, J.-M. (2015). Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. *Pattern Recognition Letters*, 66:81–90.

[202] Shi, Z. and Burr, D. (2016). Predictive coding of multisensory timing. *Current Opinion in Behavioral Sciences*, 8:200–206.

[203] Shibata, T. and Wada, K. (2011). Robot therapy: a new approach for mental healthcare of the elderly–a mini-review. *Gerontology*, 57(4):378–386.

[204] Shteynberg, G. (2010). A silent emergence of culture: The social tuning effect. *Journal of Personality and Social Psychology*, 99:683–689.

[205] Shteynberg, G. (2015). Shared attention. *Perspectives on Psychological Science*, 10(5):579–590.

[206] Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech Language*, 67:101178.

[207] Spatola, N., Belletier, C., Normand, A., Chausse, P., Monceau, S., Augustinova, M., Barra, V., Huguet, P., and Ferrand, L. (2018). Not as bad as it seems: When the presence of a threatening humanoid robot improves human performance. *Science Robotics*, 3.

[208] Stahlschmidt, S. R., Ulfenborg, B., and Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2).

[209] Staudte, M. and Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition*, 120(2):268–291.

[210] Steinbeis, N. (2016). The role of self-other distinction in understanding others' mental and emotional states: neurocognitive mechanisms in children and adults. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371.

[211] Steinhaeusser, S. C. and Lugrin, B. (2022). Effects of colored leds in robotic storytelling on storytelling experience and robot perception. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1053–1058.

[212] Stenzel, A., Chinellato, E., Bou, M. A., Ángel P. Del Pobil, Lappe, M., and Liepelt, R. (2012). When humanoid robots become human-like interaction partners: Corepresentation of robotic actions. *Journal of Experimental Psychology: Human Perception and Performance*, 38:1073–1077.

[213] Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., and Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological psychiatry*, 84(9):634–643.

[214] Stocker, A. A. and Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4):578–585.

[215] Stoffregen, T. A., Gorday, K. M., Sheng, Y.-Y., and Flynn, S. B. (1999). Perceiving affordances for another person's actions. *Journal of Experimental Psychology: Human Perception and Performance*, 25(1):120.

[216] Strait, M., Lier, F., Bernotat, J., Wachsmuth, S., Eyssel, F., Goldstone, R., and Sabanovic, S. (2020). A three-site reproduction of the joint simon effect with the nao robot. *ACM/IEEE International Conference on Human-Robot Interaction*, pages 103–110.

[217] Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., and Mittal, A. (2016). Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In Hua, G. and Jégou, H., editors, *Computer Vision – ECCV 2016 Workshops*, pages 337–348, Cham. Springer International Publishing.

[218] Sun, R., Peterson, T., and Merrill, E. (1999). A hybrid architecture for situated learning of reactive sequential decision making. *Applied Intelligence*, 11(1):109–127.

[219] Surtees, A., Apperly, I., and Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition*, 129:426–438.

[220] Surtees, A., Samson, D., and Apperly, I. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, 148:97–105.

[221] Syrdal, D. S., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2009). The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adaptive and emergent behaviour and complex systems*.

[222] Teufel, C., Alexis, D. M., Todd, H., Lawrance-Owen, A. J., Clayton, N. S., and Davis, G. (2009). Social cognition modulates the sensory coding of observed gaze direction. *Current Biology*, 19(15):1274–1277.

[223] Teufel, C., Fletcher, P. C., and Davis, G. (2010). Seeing other minds: attributed mental states influence perception. *Trends in cognitive sciences*, 14(8):376–382.

[224] Thellman, S. and Ziemke, T. (2020). Do you see what i see? tracking the perceptual beliefs of robots. *Iscience*, 23(10):101625.

[225] Thomaz, A. L., Lieven, E., Cakmak, M., Chai, J. Y., Garrod, S., Gray, W. D., Levinson, S. C., Paiva, A., and Russwinkel, N. (2019). Interaction for task instruction and learning. In *Interactive task learning: Humans, robots, and agents acquiring new tasks through natural interactions*, pages 91–110. MIT Press.

[226] Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.

[227] Tomasello, M. (2003). *The key is social cognition*, pages 47–57. MIT Press.

[228] Tomasello, M. (2014). The ultra-social animal. *European Journal of Social Psychology*, 44(3):187–194.

[229] Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28:675–691.

[230] Tsfasman, M., Philippsen, A., Mazzola, C., Thill, S., Sciutti, A., and Nagai, Y. (2022). The world seems different in a social context: A neural network analysis of human experimental data. *PLOS ONE*, 17(8):1–24.

[231] Tuncer, S., Gillet, S., and Leite, I. (2022). Robot-mediated inclusive processes in groups of children: From gaze aversion to mutual smiling gaze. *Frontiers in Robotics and AI*, 9.

[232] Turner, R. S. (1977). Hermann von helmholtz and the empiricist vision. *Journal of the History of the Behavioral Sciences*, 13(1):48–58.

[233] Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., and Baik, S. W. (2018). Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, 6:1155–1166.

[234] van Turnhout, K., Terken, J., Bakx, I., and Eggen, B. (2005). Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of the 7th International Conference on Multimodal Interfaces*, ICMI '05, page 175–182, New York, NY, USA. Association for Computing Machinery.

[235] Vannucci, F., Di Cesare, G., Rea, F., Sandini, G., and Sciutti, A. (2018). A robot with style: Can robotic attitudes influence human actions? In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 1–6.

[236] Varela, F. J., Thompson, E., and Rosch, E. (2016). *The embodied mind: Cognitive science and human experience*. MIT press.

[237] Varrasi, S., Di Nuovo, S., Conti, D., and Di Nuovo, A. (2019). Social robots as psychometric tools for cognitive assessment: a pilot test. In *Human Friendly Robotics*, pages 99–112. Springer.

[238] Vernon, D. (2014). *Artificial cognitive systems: A primer*. MIT Press.

[239] Vernon, D., von Hofsten, C., and Fadiga, L. (2016). Desiderata for developmental cognitive architectures. *Biologically Inspired Cognitive Architectures*, 18:116–127.

[240] Vertegaal, R., Van der Veer, G., and Vons, H. (2000). Effects of gaze on multiparty mediated communication. In *Graphics interface*, pages 95–102.

[241] Vinanzi, S., Cangelosi, A., and Goerick, C. (2020). The role of social cues for goal disambiguation in human-robot cooperation. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 971–977.

[242] Waldhart, J., Clodic, A., and Alami, R. (2019). Reasoning on shared visual perspective to improve route directions. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–8.

[243] Wallkötter, S., Tulli, S., Castellano, G., Paiva, A., and Chetouani, M. (2021). Explainable embodied agents through social cues: A review. *J. Hum.-Robot Interact.*, 10(3).

[244] Warren, Z. E., Zheng, Z., Swanson, A. R., Bekele, E., Zhang, L., Crittendon, J. A., Weitlauf, A. F., and Sarkar, N. (2015). Can robotic interaction improve joint attention skills? *Journal of Autism and Developmental Disorders*, 45:3726–3734.

[245] Watson, A. B. and Pelli, D. G. (1983). Quest: A bayesian adaptive psychometric method. *Perception & psychophysics*, 33(2):113–120.

[246] Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598–604.

[247] Wertheimer, M. (1938). *Laws of organization in perceptual forms*, pages 71–8. Kegan Paul, Trench, Trubner & Company, Berlin, Boston.

[248] Wiese, E., Metta, G., and Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in psychology*, 8:1663.

[249] Wiese, E., Wykowska, A., Zwickel, J., and Müller, H. J. (2012). I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PLOS ONE*, 7(9):1–7.

[250] Wood, L. J., Robins, B., Lakatos, G., Syrdal, D. S., Zaraki, A., and Dautenhahn, K. (2019). Developing a protocol and experimental setup for using a humanoid robot to assist children with autism to develop visual perspective taking skills. *Paladyn, Journal of Behavioral Robotics*, 10(1):167–179.

[251] Wykowska, A., Wiese, E., Prosser, A., and Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLOS ONE*, 9(4):1–11.

[252] Xiang, P., Yang, X., and Shi, Y. (2007). Web page segmentation based on gestalt theory. In *2007 IEEE International Conference on Multimedia and Expo*, pages 2253–2256.

[253] Yadollahi, E., Couto, M., Dillenbourg, P., and Paiva, A. (2022a). Do children adapt their perspective to a robot when they fail to complete a task? In *Interaction Design and Children*, IDC '22, page 341–351, New York, NY, USA. Association for Computing Machinery.

[254] Yadollahi, E., Couto, M., Dillenbourg, P., and Paiva, A. (2022b). Motivating children to practice perspective-taking through playing games with cozmo. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1482–1489.

[255] Ye, P., Wang, T., and Wang, F.-Y. (2018). A survey of cognitive architectures in the past 20 years. *IEEE Transactions on Cybernetics*, 48(12):3280–3290.

[256] Yonezawa, T., Yamazoe, H., Utsumi, A., and Abe, S. (2007). Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, ICMI '07, page 140–145, New York, NY, USA. Association for Computing Machinery.

[257] Zhao, X., Cusimano, C., and Malle, B. F. (2015). In search of triggering conditions for spontaneous visual perspective taking. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 2811–2816.

[258] Zhao, X., Cusimano, C., and Malle, B. F. (2016). Do people spontaneously take a robot's visual perspective? In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 335–342.