Computational Approaches to Interaction-Shaping Robotics

SARAH GILLET

Doctoral Thesis
Stockholm, Sweden 2024

Division of Robotics, Perception and Learning
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology
SE-100 44 Stockholm, Sweden

TRITA-EECS-AVL-2024:60
978-91-8106-006-5

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av Teknologie doktorsexamen i datalogi September 5, 2024, 2pm i F3, Lindstedtsvägen 26, Stockholm.

Tryck: Universitetsservice US AB

For my family,
With endless love and appreciation for your unwavering support.

Abstract

The goal of this thesis is to develop computational approaches generating autonomous social robot behaviors that can interact with multiple people and dynamically adapt to shape their interactions. Positive interactions between people impact their wellbeing and are essential to a fulfilled and healthy life. In this thesis, we coin the term Interaction-Shaping Robotics (ISR) as the study of robots that shape interactions between other agents, e.g., people, and capture previous efforts from the Human-Robot Interaction (HRI) community and emphasize the potential positive or negative, intended or unintended effects of these robots. Previous efforts have explored phenomena that indicate interaction-shaping capabilities of social robots, however, how to develop autonomous social robots that can adapt to positively shape interactions between people based on perceived human-human dynamics remains largely unexplored. In this thesis, we contribute to the technical advancement of social interaction-shaping robots by developing heuristics and machine learning methods and demonstrating their effectiveness in studies with real users. We focus on shaping behaviors, i.e., balancing people's participation in interactions to foster inclusion among newly-arrived and already present children in a music game and support adult second language learners and native speakers in a language game. Especially when leveraging learning techniques, an effective interaction-shaping robot needs to act socially appropriately. We design heuristics that are appropriate by design and establish the feasibility of autonomy for interaction-shaping robots through minimal perception of group dynamics and simple behavior rules. Allowing for learning behaviors for more complex interactions, we provide a formal definition of the problem of interaction-shaping and show that using imitation learning (IL) or offline reinforcement learning (RL) based on previously collected HRI data is feasible without compromising the interaction. To meet the challenge of acting appropriately, we explore techniques applied prior to deployment when learning offline from data and shielding - a technique from the safe RL community - to eventually allow for learning during deployment in interaction. Overall, this thesis demonstrates the feasibility and promise of computational methods for autonomous interaction-shaping robots and demonstrates that these methods generate effective and appropriate robot behavior when balancing participation to ensure the inclusion of all human group members.

Keywords

Human-robot interaction, social robotics, behavior generation, multiparty interaction, human-human dynamics, machine learning

Sammanfattning

Målet med denna avhandling är att utveckla beräkningsbaserade metoder för att generera autonoma sociala robotbeteenden som kan interagera med flera människor och dynamiskt anpassa sig för att forma deras interaktioner. Positiva interaktioner mellan människor påverkar deras välbefinnande och är avgörande för ett meningsfullt och hälsosamt liv. I denna avhandling myntar vi termen "Interaction-Shaping Robotics"(ISR) som studerandet av robotar som formar interaktioner mellan andra aktörer, t.ex. människor, och sammanställer tidigare studier inom människ-robot-interaktion (eng. Human-Robot Interaction, HRI) samt betonar den potentiella positiva eller negativa, avsiktliga eller oavsiktliga, inverkan av dessa robotar. Tidigare studier har utforskat fenomen som indikerar på interaktionsformande förmågor hos sociala robotar, men utvecklandet av autonoma sociala robotar som kan anpassa sig för att positivt forma interaktioner mellan människor baserat på observerad människa-till-människa dynamik är fortfarande till stor del outforskat. I denna avhandling bidrar vi till den tekniska utvecklingen av sociala interaktionsformande robotar genom att utveckla heuristiker och maskininlärningsmetoder och demonstrera deras effektivitet i studier med användare. Vi fokuserar på att forma beteenden, d.v.s. balansera människors deltagande i interaktioner för att främja inkludering bland nyanlända och redan närvarande barn i ett musikspel och stödja vuxna andraspråksinlärare och modersmålstalare i ett språkspel. Särskilt när man utnyttjar maskininlärningsmetoder, behöver en effektiv interaktionsformande robot agera socialt korrekt. Vi designar heuristiker som är lämpliga by design" och fastställer genomförbarheten av autonomi för interaktionsformande robotar genom minimal perception av gruppdynamik och enkla beteenderegler. Genom att tillåta inlärning av beteenden för mer komplexa interaktioner, tillhandahåller vi en formell definition av problemet av interaktionsformande och visar att användning av imitationsinlärning (eng. imitation learning, IL) off-line förstärkningsinlärning (eng. reinforcement learning, RL), baserat på tidigare insamlad HRI-data är genomförbart utan att kompromissa med interaktionen. För att möta utmaningen att agera korrekt, utforskar vi tekniker som tillämpas innan implementering när man lär sig off-line från data och ”shielding” - en teknik inom säker RL - för att så småningom möjliggöra inlärning under implementering vid interaktion. Sammanfattningsvis visar denna avhandling genomförbarheten och utsikten av beräkningsbaserade metoder för autonoma interaktionsformande robotar och demonstrerar att dessa metoder genererar effektiva och lämpliga robotbeteenden när de balanserar deltagande för att säkerställa inkludering av alla mänskliga gruppmedlemmar.

Acknowledgements

It is summer, I am sitting outside, there is a warm breeze going, and I come to realize that one important learning of my Ph.D. journey was to be appreciative and grateful for people spending their valuable time to chat about life and career, read drafts and listen to presentations, or glance at what I send for feedback.

First, I would like to thank Iolanda, my wonderful supervisor. Your kindness, patience, and honest caring, on the one hand, and your knowledge and guidance through HRI go way beyond what I could have asked for. I am also very grateful for the time I could spend at Yale University with Marynel and the meetings, guidance, and advice I received from you long before and long after my visit. Your approach to professorship, lab management, and technical guidance is truly inspiring, and I am grateful for working closely with you. I was fortunate to find very dear friends during this journey. Sanne and Irmak, I am very happy to have spent most of this rollercoaster ride with you two, through paper acceptance and rejection, traveling to conferences or for research visits, challenges when recruiting participants and running experiments, and this thing they called a pandemic. Katie and Ilaria, we met just before or right during the pandemic. I am thinking back to many wonderful hours in which I learned about pizza, sourdough, pasta, feminism, and British culture and accents. I am grateful for your friendship but also for your advice and honest opinion on career, approaches to academic life, and the struggles and challenges of the Ph.D.

I have been fortunate to have many awesome collaborators. First and foremost, I want to thank Ronald and Teresa. Working and collaborating with you not only produced awesome papers and interesting research outcomes but also let me reflect on my communication and approach to work, which I found invaluable to me and my journey ahead. I am also very grateful to have worked with Sydney, Kate, and the Yale IMG group. I experienced many inspirational hours collaborating with Hadas, Nicole, Marynel, Sarah, and Scaz, which shaped how I think and look at collaborations. I also want to thank Sarah Sebo for embarking with me on the journey with Marynel, Sean, and Iolanda, who, in the end, defined interaction-shaping robotics and gave my research a new home. Thank you, Giulia, for joining and pushing for our teenage dream. Special thanks also go to Alex; during all our collaborations, I valued your patience, sense of people, and knowledge about robots a lot, and I am grateful for every fun and also intense hour we spent together.

I also wanted to thank everyone who was and is in Iolanda's group: Daniel, Simon, Gaspar, Youssef, Hannah, William, Amandus, Sylvaine, Elmira, Arzu, Rebecca, Katie, Ilaria, and those who visited, Sofia and Merle. Thank you for the many fun lunches, the fun times, and the interesting discussions in the group meetings. Special thanks to Joe, who was brave enough to share an office with me during the most intense time of finishing this thesis and from whom I learned much about designerly ways of knowing in HRI.

I also want to express a special thanks to Patric for making the experience in this division unique, leaving no doubt that every student is cared for, to Jana and Atsuto for their time to listen and for valuable advice during the mentoring meetings, and to Hedvig for being my additional supervisor and for showcasing how to balance academia and the love for music and orchestra play.

I would also like to thank doctoral education support for their help during my Ph.D. and especially toward the finalization and defense, Gabriel Skantze for serving as the advanced reviewer, Tony Belpaeme for agreeing to be the opponent, and Valentina Fantasia, Goren Gordon, Malte Jung, Catharine Oertel, and Anara Sandygulova for serving on my grading committee.

Mein größter Dank gebührt meinen Eltern Gabi und Klaus, meinem Partner Simon und meiner Schwester Jana mit Partner Damian. Ohne Eure uneingeschränkte Liebe und Unterstützung wäre ich nicht mutig genug gewesen, um zunächst einmal mein Informatikstudium zu beenden und dann einen unbefristeten Job aufzugeben, um meinem Traum einer Forschungslaufbahn nachzugehen.

The Ph.D. is only one part of the academic journey, and the defense may be a waypoint before the next part begins. I hope everyone I was fortunate to meet so far will be part of my journey ahead.

Sarah Gillet
Stockholm, June 2024

List of publications

This thesis is based on the following publications:

A: **Sarah Gillet**, Marynel Vázquez, Sean Andrist, Iolanda Leite, Sarah Sebo. *Interaction-shaping Robotics: Robots That Influence Interactions between Other Agents.* In ACM Transactions on Human-Robot Interaction, Vol. 13, No. 1, 2024

B: **Sarah Gillet***, Ronald Cumbal*, André Pereira, José Lopes, Olov Engwall, Iolanda Leite. *Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels.* In Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, 2021, New York, NY, USA. (HRI '21)

C: **Sarah Gillet**, Wouter van den Bos, Iolanda Leite. *A social robot mediator to foster collaboration and inclusion among children.* In Proceedings of Robotics: Science and Systems, 2020, Corvalis, Oregon, USA. (RSS '20)

D: **Sarah Gillet***, Katie Winkle*, Giulia Belgiovine*, Iolanda Leite. *Ice-Breakers, Turn-Takers and Fun-Makers: Exploring Robots for Groups with Teenagers.* In Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication, 2022, Naples, Italy. (RO-MAN 2022)

E: **Sarah Gillet**, Daniel Marta, Mohammed Akif, Iolanda Leite. *Shielding for socially appropriate robot listening behaviors.* In Proceedings of the 33rd IEEE International Conference on Robot and Human Interactive Communication, 2024, Pasadena, USA. (RO-MAN 2024)

F: **Sarah Gillet**, Maria Teresa Parreira, Marynel Vázquez, Iolanda Leite. *Learning Gaze Behaviors for Balancing Participation in Group Human-Robot Interactions.* In Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction, 2022, New York, NY, USA. (HRI '22)

*The authors agreed to share first authorship.

The author has contributed to the following articles as well, although they are not included in this thesis:

X-1: **Sarah Gillet**. *A computational lens to interaction-shaping robotics.* In Pioneers Workshop of Robotics: Science and Systems, 2024

X-2: Georgios Hadjiantonis, **Sarah Gillet**, Marynel Vázquez, Iolanda Leite, Fethiye Irmak Doğan. *Let's move on: Topic Change in Robot-Facilitated Group Discussions* . In Proceedings of the 33rd IEEE International Conference on Robot and Human Interactive Communication, 2024, Pasadena, USA. (RO-MAN 2024)

X-3: Hadas Erel*, Marynel Vázquez, Sarah Sebo, Nicole Salomons, **Sarah Gillet**, Brian Scassellati. *RoSI: A Model for Predicting Robot Social Influence.* In ACM Transactions on Human-Robot Interaction, Vol. 13, No. 2, 2024

X-4: Maria Teresa Parreira, **Sarah Gillet**, Iolanda Leite. 2023. *Robot Duck Debugging: Can Attentive Listening Improve Problem Solving* In Proceedings of the 25th International Conference on Multimodal Interaction, New York, NY, USA. (ICMI '23).

X-5: Maria Teresa Parreira, **Sarah Gillet**, Katie Winkle, Iolanda Leite. *How Did We Miss This? A Case Study on Unintended Biases in Robot Social Behavior*. In Companion of the 18th ACM/IEEE International Conference on Human-Robot Interaction, 2023, New York, NY, USA. (HRI '23)

X-6: Kate Candon, Helen Zhou, **Sarah Gillet**, Marynel Vázquez. *Verbally Soliciting Human Feedback in Continuous Human-Robot Collaboration: Effects of the Framing and Timing of Reminders.* In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)

X-7: Sylvaine Tuncer, **Sarah Gillet**, Elizabeth J Carter, Iolanda Leite. *Robot-Mediated Inclusive Processes in Groups of Children: From Gaze Aversion to Mutual Smiling Gaze.* In Frontiers in Robotics and AI, Volume 9, 2022.

X-8: **Sarah Gillet**. *Autonomous Robot Behaviors for Shaping Group Dynamics.* In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI Pioneers), 2021

X-9: Fethiye Irmak Doğan, **Sarah Gillet**, Elizabeth J. Carter, Iolanda Leite. *The impact of adding perspective-taking to spatial referencing during human–robot interaction.* In Robotics and Autonomous Systems, Volume 134, 2020

*All authors contributed equally to this research.

List of abbreviations

BC Backchannels

BOP Backchanneling Opportunity Points

HRI Human-Robot Interaction

IL Imitation Learning

IPO Input-Process-Output

ISR Interaction-Shaping Robotics

LSTM Long-Short Term Memory

MDP Markov Decision Process

RL Reinforcement Learning

WIS Weighted Importance Sampling

Contents

Part I

Overview

Chapter 1

Introduction

”I can’t interrupt them and then I am screwed. So I wait and wait and wait. So this microphone [robot] will be really good so I could talk to the others“ (Participant W2, 13 years old, upon reflecting on how robots from the literature could help them interact with other teenagers, Paper D).

Humans are "[...] by nature [..] social animal[s] [...]" (Aristotle, Politics) who thrive through social interactions, relationships, and communication. This hypothesis by Aristotle was heavily debated through the centuries, but more recent research found that positive interactions between people contribute to feelings of happiness, security, self-esteem, and pleasure [165, 84], and close and positive relationships between people were shown essential to living a fulfilled and healthy life [207, 214]. In cases where relationships between people are negative or nonexistent, people suffer from social rejection, loneliness, and poorer physical health [93, 191].

We could argue that sustaining positive interactions should be a solely human task, but striving for positive relationships might be harder for some than others. For example, people might have to overcome barriers, e.g., relating to inclusion, power, or skill, that hinder positive human-human interactions, or they may have to learn effective interaction skills like the teenager in the opening quote above. Therefore, researchers explore tools, including robots, to influence and enhance interactions and relationships between people [151, 91] and society as a whole [185]. For example, research explores supporting participation during educational collaborations [192] through a technique called collaboration scripts [213] or through digital awareness tools calling attention to the distribution of speech amounts [102].

Focusing on robots as the supporting tool, pioneering prior work from the field of HRI provides initial evidence that robots can influence people beyond the reciprocal interaction between a person and a robot. **Robots can influence the interactions between people.** We shall refer to these robots as interaction-shaping robots (Paper A). Robots have been found to improve situations of conflicts [105] by inviting children to more constructive conflict resolution [184], shape participation in conversations [140, 139] and discussions [199], and support the process of inclusion in social [143, 142] or work-related environments [194]. These and other prior works have primarily centered on uncovering the potential of interaction-shaping robots within distinct contexts and studying empirical, psychological, and cultural aspects unique to human-human and human-robot interactions.

In this thesis, **robots are viewed as a tool or technology** that we eventually want to deploy in the real world. To be effective in real-world interactions, robots will first need to perceive these human-human interactions and their dynamics. For example, the robot might need to understand if there is a conflict and what the source of the conflict is. It might need to perceive if everyone contributes to the discussion or if someone is dominating it. Works from affective computing but also HRI, Human-Computer Interaction, and Computer-Supported Collaborative Work explored how to perceive groups and their dynamics [68], e.g., cohesion [182], dominance [166], or engagement [119]. Secondly, robots will need to act considering the perceived people and their interactions to have the intended positive impact on human-human interactions. The HRI community has explored social behaviors that act on the perception of their interaction partner through heuristics [199], Imitation Learning (IL) [101], or Reinforcement Learning (RL) [3] exploring the generation of learning curricula [157], sustaining engagement [168], positioning in groups [212] or adapting to humor preferences [216]. However, how to close the loop between perceiving human-human interactions and acting to shape these interactions remains largely unexplored.

This thesis contributes **computational approaches to the generation of autonomous social robot behaviors** that can interact with **multiple people** and dynamically **adapt to human-human dynamics**, positively shaping interactions between people. The goal of the proposed computational approaches, i.e., heuristics or machine learning methods, is to close the loop between the perceived dynamics of human-human interactions and effective social robot behaviors to shape interactions between people. In this sense, this thesis studies generating behavior as repeatedly performing two steps: 1) sense people and their interactions, and 2) reason and act upon the observation to positively shape the human-human interaction. Based on this goal, this thesis provides contributions toward the following two research questions:

RQ1 Can autonomous social robots capable of adapting to dynamic groups effectively shape interactions?

RQ2 How can we leverage machine learning methods to learn autonomous social robot behaviors that aim to shape interactions?

To answer the first research question, we first focus on a specific interaction context in which we expect imbalances to occur that the robot can then mediate - an interaction between a language learner and a native speaker in a language game. By leveraging results from previous HRI but also human-human experiments, we build a dynamically adapting gaze policy that effectively shapes the interaction between the learner and native speaker, leading to more balanced interactions. Our results show promise for autonomous interaction-shaping robots for groups of adults. When we consider groups of children, a robot's support might be especially important when they have newly arrived in a country without command of the language. We explored how a robot could foster inclusion and collaboration among newly arrived and already present children using a language-free music-mixing game. Our results suggest that the robot was able to encourage newly arrived children to be more outgoing and encourage the other children to collaborate more and act more prosocially when voluntarily giving away stickers. With the goal of moving toward more complex, real-world interactions, this thesis then uses co-design methods to explore how teenagers imagine, design, and use a robot to support their group discussions. Our results show that teenagers use the robot in distinct and fluent ways, which reflects their development and interactions as a group, further highlighting the need for robots that perceive and adapt to the dynamics of group interactions.

Exploring the second research question, this thesis proposes an approach to learning online in interactions despite the need for exploration. Exploration, necessary when learning online, might lead to inappropriate behaviors. Therefore, we explore the use of shields - a method from the safe RL community - to guide the robot when learning online. To create the shields, we rely on a human-human dataset and a notion of appropriateness that does not indicate an immediate need to act. Therefore, this thesis proposes to use offline methods based on HRI datasets. We use offline RL or IL to learn behaviors that use gaze to balance participation. As offline methods allow us to inspect the behavior before deployment, we analyze the learned behavior policies for their appropriateness. Even though the learned policies did not improve over the heuristic baselines, they show that learning approaches for interaction-shaping robotics might be feasible without compromising the interaction.

In summary, this thesis contributes **computational approaches** in the form of **heuristics** and **machine learning methods** to shape interactions between humans, specifically when **balancing participation**. By studying the contributed approaches in **HRI** with human participants, this thesis further contributes a deeper understanding on how robots can balance participation, i.e., to foster inclusion or encourage equal contributions during language learning. Adding a theoretical perspective, this work also coins the term Interaction-Shaping Robotics as the study of robots that shape interactions between other agents, shedding light not only on computational challenges but also methodological and ethical challenges. As such, this thesis contributes to the technical, user studies, and theoretical perspectives of HRI to develop robots that encourage and support the flourishing of human-human interactions.

1.1 Thesis outline

The following chapters will first provide the background literature covering two perspectives. First, aspects of human communication and group dynamics will be explained before moving on to the computational perspective describing the foundations of reinforcement and imitation learning (Chapter 2). Then, Chapter 3 will provide an overview of related work concerning multiparty HRI, robots interacting in and influencing groups of people, and computational approaches to generating robot behavior. This thesis' contributions will be summarized and briefly detailed in Chapter 4. Lastly, Chapter 5 will discuss the findings and implications of the contributions made in this thesis, interaction-shaping robotics in general, and respective limitations. Chapter 6 will conclude this thesis.

Chapter 2

Background

Social robots are at the core of this thesis and researchers have put forward a variety of definitions that explain what a social robot is [15, 27, 58, 88]. This thesis agrees with the definition by Bartneck and Forlizzi [15], which refers to social robots as an "autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact." Social robots communicating with people and following behavioral norms is certainly the goal of this research; however, in the process of researching methods and techniques that fulfill this goal, we lower the assumption on the definition by following the definition by [88] that defines a social robot as a physically embodied robot that has a social interface to interact and communicate with people.

The remainder of this chapter will first discuss in Section 2.1 how humans communicate, specifically, non-verbal elements of speech, such as eye gaze and backchanneling. Then, Section 2.2 provides a brief overview of research concerning human groups and which factors might influence the group dynamics. Lastly, techniques that generate behavior, such as heuristics, reinforcement learning, and imitation learning, are discussed in Section 2.3.

2.1 Communicative behavior

Robots can use different communication modalities and forms to interact with people. This section briefly reviews communication in human-human interactions, focusing on forms of communication relevant to the work in this thesis, i.e., eye gaze and backchanneling behavior.

Eye gaze

People are generally sensitive to others' eye gaze. For example, seeing someone's eyes gazing in one direction leads to a rapid attention shift toward the direction of the gaze, even if the eyes are in a still photograph [115, 64]. This reflex has further been found to be hard to control. When participants were told to look in the opposite direction, they still first reacted to looking in the direction of the gaze before gazing at the requested location with a delay [48, 47, 63].

People use gaze and the observation of other people's gaze in a variety of communicative behaviors such as when coordinating conversations [109], speech [11] or attention [64]. In conversations, gaze supports turn-taking and manages who holds the floor [108]. The speaker typically looks away at the beginning of an utterance to claim the turn, and looking away during a pause indicates that the speaker intends to hold the floor [108]. To coordinate a turn exchange, the first speaker typically looks toward the other person and establishes a momentary mutual gaze before the other then averts their gaze to begin their turn [145].

When people are gazed at, they generally feel attended to but might also feel uncomfortable from being observed [11]. Research further found that being gazed at improves learning results in college students [30] and even increases the recall when the gaze was directed through a camera in a video conversation [66].

Part of behaviors involving gazing at people is averting the gaze. For example, gaze aversion has been found to imply cognitive processing in conversations [72] and, when used by a speaker, could imply the speaker is deeply thinking about the words being used [11].

Backchanneling

One aspect of human-human communication is actively listening while the other talks. One aspect of listening behavior are Backchannels (BC)– short vocal or non-vocal expressions not interrupting the current speaker's turn [80]. BCs are portrayed through body language, such as head nods, smiles, and short vocalizations (*'uh-huh', 'hmm'*). BCs play an important conversational role, by signaling attentiveness or emotion to a speaker. Backchanneling has also been found to impact the perceived personality and rapport building [24, 96].

Backchanneling occurs in Backchanneling Opportunity Points (BOP) [75], sometimes also called backchanneling relevant spaces. Previous work indicates that, on average, there are 3.5 times more BOPs than actual backchannels [89]. In addition, recent research has shown that backchanneling behavior is idiosyncratic [25], i.e., it is peculiar to an individual, differing significantly between people.

This recent finding is in line with previous work, which showed that BC behavior is correlated with the personality of the listener [95], and adjacent factors such as status or dominance also influence the amount of backchanneling [171, 133]. Further, participants identifying as female backchanneled more than participants identifying as male [20, 171, 172]. Interestingly, female and male observers might also interpret the function of backchannels in different ways [135]. In the study's specific context, females tended to associate backchannels with interest, whereas males interpreted them as a sign of uncertainty.

2.2 Groups and group dynamics

This thesis views groups as “two or more individuals who are connected by and within social relationships”[59] and “influence one another in such a manner that each person influences and is influenced by each other person”[183]. Whereas this thesis defines two people as the minimal number of human members of a group, it acknowledges the robot as a third group member. The minimal number of three agents in a group aligns with other research that argues for a minimal number of three for forming a group [217]. Overall, this thesis often refers to human-human interactions and dynamics instead of groups to differentiate between the robot’s role in the group and its goal - to shape the interaction between the human group members.

Groups may engage in conversations in which group members may be participants [37]. Participants in a conversation can take different roles [73]. Essential for each conversation are the “speaker” and “addressee” [37]. Other “ratified participants” [73] are called “side participants” in a conversation. For example, other group members generally part in the interaction but not in the current conversational exchange can be called side participants. In addition, the literature considers overhearers and differentiates between two types of overhearers. Bystanders [37, 73] are “non-participants” [37] acknowledged by the speaker but do not take part in the conversations. Eavesdroppers are all other listeners, e.g., out of sight or otherwise not acknowledged by the speaker [73]. The different roles people take and how they shift during a conversation are important for understanding spoken interactions [82]. Further, the different listener roles influence how the speaker gazes toward them [140], e.g., gazing more at the addressee than a side participant.

In addition, an important distinction is often made between groups and teams. While groups are a broader term with a minimal definition as above, teams are groups that share a goal or task and work together to produce a product, a service, or a decision [81, 76]. The success, i.e., team performance, is, according to the Input-Process-Output (IPO) Model, dependent on inputs and processes. Inputs are defined by individual-level factors such as personality, team-level factors, i.e., diversity and group size, and contextual and environmental factors. Processes are concerned with sustaining interdependence, coordinating the interaction, setting and monitoring goals, adapting structures, maintaining cohesion, and resolving conflicts. In addition to performance, the IPO also considers outputs on the team level, such as improved procedures and member-level outcomes, e.g., satisfaction or personal development [59].

As the definition above indicates, group phenomena that emerge when groups interact are subject to group social influence. For example *conformity* [43] means agreeing to the majority view in a group. *Minority influence* discusses the power of the few[132]. *Power and social status* concern the difference between members in their capacity to influence the others [62, 18]. *Leadership* studies the guidance of group members to achieve individual or collective goals [94].

Group phenomena also relate to feelings of inclusion and identity. Specifically, group researchers argue that humans have a need to belong [16] and that group members help each other to avoid social and emotional loneliness. When group members get deliberately excluded from a group, this process is referred to as *ostracism*. Researchers found that ostracism is stressful for people and results in anger and sadness [218]. On the other hand, college students reported being less lonely if they were part of a cohesive group [12], and generally, people were found to be healthier and happier if they belonged to groups [83], e.g, service organizations or social clubs, with an increase in effect if people belonged to multiple groups [160].

A phenomenon that does not necessarily arise from but might influence small group interactions is *ingroup-favoritism*. Ingroup-favoritism has been phrased by the findings that people tend to act more favorably toward ingroup members than outgroup members [29, 198]. The formation of groups inevitably leads to processes between groups - intergroup processes. When intergroup effects concern biases, these effects are described as the natural favoring of one's own group (ingroup) over other groups (outgroup). This ingroup-outgroup bias then affects the perception of and behaviors toward the outgroup [59] with people acting more favorably toward the ingroup than the outgroup [29, 198]. To overcome those ingroup-outgroup biases, the literature proposes the contact hypothesis that suggests that relationships between groups can be improved through positive contact in a joint interaction [6, 220].

To study groups and their dynamics, social science literature frequently focuses on [1]: *Ingroup identification*, the individuals' perception of themselves as members of the group [13, 117]; *cohesion*, the inside perspective on the forces that keeps the group as a group [44]; and *entitativity*, the outside perception of the groupness of a social group [32].

This thesis builds upon those prior works, laying the foundation for understanding and studying groups by drawing from the described phenomena, e.g., intergroup processes, to understand the specific dynamics of the context the robot is deployed in and focus the robot's perception and action system on aspects of the group that seem relevant to the phenomena and measurable dynamics.

2.3 Methods for generating social robot behavior

This thesis contributes heuristics and machine learning methods to generate social robot behavior. Therefore, this section defines terminology around heuristics and presents the foundations for Reinforcement Learning (RL) and Imitation Learning (IL).

Heuristics

The term *heuristic*, as used in the HRI community, typically refers to a method or algorithm that generates social robot behavior based on a set of rules, e.g., when the person who we assigned to the outgroup contributes verbally, praise this person

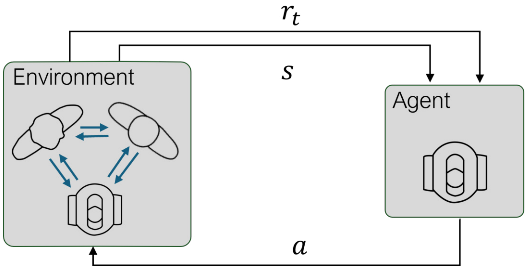


Figure 2.1: Visualization of the general RL problem adapted to the context of this thesis where the agent is a social robot and the environment is characterized by the interactions between people and the robot.

[194]. These heuristics are typically hand-crafted based on expert knowledge [193] or result from observing and analyzing data from human-human interactions [215].

Reinforcement Learning (RL)

RL provides a framework for solving sequential decision-making problems in which actions influence the environment in the immediate moment and subsequent moments typically formalized as a Markov Decision Process (MDP). In the following, this thesis will explain the formalization of RL as put forward by Sutton and Barto [196]. As visualized in Figure 2.1, the MDP is formed by the *agent*, the learner and decision maker, and the *environment* it interacts with. In addition, the environment provides a reward corresponding to the value the agent wants to maximize over the time it interacts with the environment. The agent might influence the reward it receives through the action it takes.

The agent perceives the environment in each discrete time step, $t = 0, 1, 2..$, as state $s_t \in \mathcal{S}$ to select the best action $a_t \in \mathcal{A}$. Once the action is executed, the altered environment provides the next state s_{t+1} and a reward $r_{t+1} \in R$. The environment's dynamics further describe a state-transition function $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ as $p(s'|s, a)$ which describes the probability that the environment is in state s' at time step $t + 1$ if it was in state s in time step t and the robot executed action a. In most social HRI contexts, $p(s'|s, a)$ is treated as unknown due to the complexity of human-human interactions, and model-free RL methods are used.

To select the best action, the agent uses a *policy* $\pi(a|s)$, which provides the probability that the current optimal agent would take action a in state s. Many RL algorithms are value-based, which means they approximate a value function $v_\pi(s)$ or Q-function $q_\pi(s, a)$ when acting according to policy $\pi(a|s)$. The value and Q function capture the expected discounted sum of future rewards the agent can expect. The discount factor γ defines the importance of future rewards. If $\gamma = 0$, the agent is considered myopic and only concerned with maximizing the immediate rewards. The Q-function is then defined as $q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^t r_{t+k+1} \mid s_t = s, a_t = a\right]$.

RL algorithms change the policy π based on the experience the agent collects in the environment to eventually provide a policy π that reflects optimal behavior, i.e., maximizes future rewards. As a consequence, the agent has to collect experiences in its environment. One challenge of RL is the trade-off between exploration and exploitation. The agent needs to exploit the knowledge it already has to collect rewards, however, it also needs to explore to find potential higher reward. Overall, an RL agent cannot succeed at the task without doing both - exploration and exploitation.

When the RL algorithms exploit the collected knowledge, we can assume that the policy π will always take the optimal action; hence, the Bellman equation describing the optimal q function is defined as $q^* = \mathbb{E}[r_{t+1} + \gamma \ \mathtt{max} q^*(s_{t+1}, a')|s_t = s, a_t = a]$.

To avoid catastrophic failures during learning, research has suggested to use safe RL methods [67] or offline RL [114].

Research interested in offline RL or batch RL comprises RL algorithms and techniques that learn a policy π from an existing batch of data without the need to collect new data for training. One challenge of offline RL is to ensure that the policy stays within the state and actions explored in the dataset and avoids visiting unseen state action pairs for which the optimal behavior cannot be defined from the dataset. One way to meet this challenge is to restrict the update via the Bellmann equation to only consider state-action pairs in the dataset $\mathcal{B}$: $q^*_{\mathcal{B}}(s_t, a_t) = \mathbb{E}[r_{t+1} + \gamma \underset{a' s.t. (s',a') \in \mathcal{B}}{\mathtt{max}} q^*(s', a')|s_t = s, a_t = a]$ [65].

Safe RL works consider methods that ensure system performance while respecting safety constraints during learning and/or deployment. One approach to safe RL is to modify the optimality criterion by adding a safety factor. Other methods alter the exploration process [67], for example, through shielding. Methods that use shielding restrict the action space A to a safe set $\mathcal{A}_{\text{safe}}(s) := \Psi(s) \subseteq \mathcal{A}$ either by providing only the list of safe actions to the robot before selecting the optimal action or by stopping the execution of unsafe actions not in the safe action set $\mathcal{A}_{\text{safe}}(s)$.

Imitation Learning (IL)

IL comprises techniques and algorithms that mimic human behavior in a task previously demonstrated by the human. IL uses the same definition of the MDP above without the reward r_t. Formally, the goal of IL is then to learn a policy $\pi(s, a)$ that follows these demonstrations as closely as possible by mapping observations to actions through a function f defined as $f(s) \rightarrow a$ with $s \in \mathcal{S}$ and $a \in \mathcal{A}$. This bare form of IL is also known as behavioral cloning. To learn the direct mapping as function f, techniques used for classification tasks can be used for discrete action spaces or regression for continuous action spaces.

Chapter 3

Related work

The work in this thesis builds upon two research areas: 1) multiparty human-robot interaction, i.e., robots interacting in groups and their effects on human-human dynamics, and 2) computational approaches to generating social robot behavior. In the following, related work from these two areas is discussed and briefly related to the contributions of this thesis.

3.1 Multiparty human-robot interaction and effects on human-human dynamics

Within recent years, the study of multiparty human-robot interaction has been established as an essential area in HRI [103, 180, 178]. Works in this area have considered robots encountering multiple people in office buildings [26, 86], robots in public places such as museums [202, 144, 55, 189], train stations [87], airports [205], or shopping centers [60, 137] as well as in hospitals [138], hotels [36], schools or care centers for children [106, 107, 221, 118]. Of specific interest to this thesis are works that consider collocated small group interactions which comprise group interactions that occur in the same physical space often focused on a shared task, for example, in collaborative tasks [105, 40, 104], dancing [98] or learning [50, 122, 174, 79, 131, 173] together, in multiparty games [211, 209, 186, 42, 148], or engaging through conversations [140, 129, 210, 204, 163].

Related work that studies the effects of robots being part of the ingroup or outgroup [111] is important when studying collocated groups with the robot being introduced as a group member or facilitator (see Table 3.1). For example, introducing the robot as ingroup versus outgroup increased the positive perception of the robot [112] and can lead to developing ingroup favoritism for robots over humans as people prefer ingroup robots over outgroup humans [61]. Social categorization has been found as one factor that influences the perception of the robot as ingroup [56].

Specifically, when humans and robots partner in teams against another human-robot team, robot behaviors affect how people perceive and behave towards robots.

Socio-emotional support and gaze behaviors exhibited by people were shown to depend on the orientation goals of a robot (competitive versus cooperative) in a game [150]. Further, the levels of warmth and comfort a robot might express can influence feelings, perceptions, and future intent to work with the robot [149]. In contexts in which group identification, group trust, and likeability of the robot are important, it was found that a robot expressing group-based emotions is favorable [41, 8].

Robots shaping human-human interactions

In [71] (Paper A), we discuss five key factors we observed as unique in the literature covering robots that shape interactions between other agents: 1) the role of the robot in the group, 2) the robot-shaping outcome, 3) the form of robot influence, 4) the type of robot communication, and 5) the timeline of robot influence on the interaction(s). Table 3.1 describes these key factors and the underlying categories. The related work below is discussed along two factors most relevant for this thesis – type of robot communication and form of robot influence – resulting in a 2x2 consideration of works with the two categories each, i.e., explicit influence and verbal or non-verbal communication, and implicit influence and verbal or non-verbal communication.

Explicit influence

When a robot shapes interactions explicitly through verbal communication, it often uses language to directly comment on aspects of the group interaction it wants to shape. For example, robots have been used as mediators in conflict situations. The robot could promote more constructive conflict-solving behavior in case of object possession conflicts among children by explicitly guiding the children through resolution strategies [184]. In another example, a robot called out inappropriate behavior, i.e., interpersonal violations, displayed by a confederate toward one of the other group members [105]. This explicit calling on the conflict led to a heightened perception of the conflict, which offered opportunities to resolve the conflict. Shamekhi et al. [181] showed how a robot could act as a facilitator and improve human-human meetings. As one aspect of meeting facilitation, they explored asking the passive participants about their opinions before accepting the group's decision.

Other opportunities for explicit influence through verbal means are given when robots act as facilitators in group interactions with the purpose of building positive interactions such as peer support groups [21] or therapy for romantic couples [206].

Combining verbal and non-verbal multimodal behaviors, robots were found effective when balancing participation through calling out the name, using lights and movement with a directive strategy among mixed groups of normal-sighted and visually impaired children [141]. However, children felt more heard by their peers when the robot used a more organic way of encouraging participation.

ISR Factor	Category	Description
Role of the Robot	Guiding Facilitator	The robot leads and directly mediates the interaction between the agents.
	Peripheral Facilitator	The robot is present and active, but is not directly involved in the interaction.
	Peer Group Member	The robot acts as a peer relative to the agents.
	Specialized Group Member	The robot adopts a special role as a group member relative to the agents.
Robot-Shaping Outcome	Cognitive	The shaping outcome is measurable in changes in cognitive attitudes and thoughts (e.g., interpersonal evaluation, feelings, intentions).
	Behavioral	The shaping outcome is measurable as a change in behavior (e.g., spatial repositioning, amount of speaking, gazing).
Form of Robot Influence	Explicit Robot Influence	The robot addresses aspects of the interaction explicitly through clear and exact communication, directly prompting or requesting a change in the interaction (e.g. calls a conflict out and asks for resolving it).
	Implicit Robot Influence	The robot implicitly addresses aspects of the interaction that could lead to a change in the interaction among the other agents.
Type of Robot Communication	Verbal	The robot uses verbal natural language to shape the interaction.
	Non-Verbal	The robot uses non-verbal behavior (e.g., gestures, gaze, movement, resource distribution) to shape the interaction.
Timeline of Robot Influence	Immediate Influence	The robot's behavior immediately shapes the interaction between the agents.
	Long-Lasting Influence	The robot's behavior shapes the interaction between the agents after the robot's interaction-shaping behavior has concluded (e.g., the following day).

Table 3.1: Key factors of interaction-shaping robots that distinctly identify mechanisms that allow robots to shape interactions. See Section 3.1 for examples.

Non-verbal behaviors might generally be understood as implicit. However, the robot Micbot, a microphone that turned to a participant to encourage speaking, could be seen as explicit prompting [199]. The study showed that turning toward the least active every now and then can balance. As such, this thesis builds upon this pioneering work by using the concept of 'least active' and encouraging them to participate through different robot behaviors, i.e., non-verbal prompts through gaze (Paper B) and whole body movement (Paper C).

Implicit influence

Related work has shown that robot behaviors might implicitly influence how other group members interact with each other. For example, making vulnerable expressions, e.g., "Sorry guys, I made the mistake this round", a robot was found to cause a ripple effect in a team of three humans and the robot that led to more trust-related behaviors [195] and improved conversational dynamics [204] within the group. Further work explored how the robot as a fourth participant might improve conversation dynamics by harmonizing engagement density through actively intervening in the conversation [126]. Among children, the reliability of a robot's task-related verbal statements was found to, in turn, influence social dynamics among children [35], i.e., children had more task-related interactions when the robot was unreliable. Interestingly, a robot that aimed to improve the social dynamics through relation-reinforcing utterances influenced the perception of team performance when compared to task-reinforcing utterances.

A larger body of work studied the implicit influence exhibited through non-verbal behaviors. For example, a non-anthropomorphic robot [90] used as a peripheral facilitator in a human-human interaction could improve interpersonal interactions and emotional support [167, 53]. Further, a robot's gaze behaviors were found to influence people's conversational roles [140, 139] as main addressee and bystander and influence who takes the turn [190, 187].

In a different line of work, Claure et al. explored the effect of a robot's fair or unfair allocation of blocks between human group members in a block stacking task [104]. The results indicate that generally perceived fairness is not influenced by the robot's unfair, performance-based allocation behavior [38] but influences the perception of dominance between group members [39]

Whereas previously mentioned related work focused on reporting positive effects, there have also been efforts to reveal which negative effects robots could have on interactions. For example, in a scenario where only one team member could ask a robot for information, that team member experienced a greater sense of exclusion from the human members of their group [194]. If the robot then supported this outgroup member through verbal encouragement or praise, their verbal participation increased. However, the robot's verbal support for the outgroup member decreased otherwise important backchannel feedback from the other human group members [179]. Similarly, in a group with a parent, a toddler, and a robot, more robot scaffolding led to less parent scaffolding, i.e., the interaction between parent and toddler was reduced [78].

Another line of work explored how two non-anthropomorphic robots excluding a participant in a ball tossing game could induce feelings of ostracism in people [52] and influence how participants behaved in a subsequent human-human interaction [51]. Participants were found to sit closer to the participant, implying a greater need to belong, after they were excluded by the robots.

3.2 Computational approaches to generating social robot behaviors

Computational HRI [200] as a subfield of HRI explores perceiving humans, their activities, intentions, and behaviors, generating communicative behaviors and affect and emotion in robots or navigation behaviors. In addition, computational HRI is concerned with learning efficiently from human teachers. The work in this thesis could be understood as generating communicative behaviors. Therefore, the following discussion of related work will mostly include works on generating communicative social robot behavior. In addition, the discussion focuses on works that use imitation learning or reinforcement learning or generate social behavior that affects, in turn, human behavior.

Imitation learning in HRI

Imitation learning, sometimes referred to as learning from demonstration in robotics [10, 19], has been used to generate robot behaviors for various human-robot interaction contexts. For example, IL has been used to learn robot behaviors from expert human demonstrations, such as through kinesthetic teaching of manipulation skills [4, 134] or modeling dyadic human-robot interactions for robot navigation [177, 197].

When generating communicative behaviors, IL approaches often use human-human interaction datasets [152]. For example, crowd-sourcing in online games or virtual worlds has been used to train behaviors for virtual agents or robots in social settings [155, 154, 28, 146]. Using specifically collected human-human interaction datasets, related work explored the generation of behavior for shopkeeper robots [45] or travel agents [46], e.g., for deciding when to take the turn [188, 49].

IL is commonly used to generate non-verbal behavior, for example, in tutoring interactions [2]. Other related work focused on generating gestures [113, 153] or non-verbal listening feedback, i.e., backchannels [147]. Deep learning techniques have been used to generate listening behaviors, e.g., with the help of Recurrent Neural Network (RNN) models, which capture the temporal dependencies of continuous signals [158]. Prior work explored Long-Short Term Memory (LSTM) layers with audio signals and word history [175] or used LSTM and multimodal input [136] combined with data augmentation method to improve backchannel predictions.

Reinforcement learning in HRI

Significant research on RL was applied to HRI [3, 123]. Related work explored different RL techniques such as multi-armed bandits [170, 169, 121] or value-based approaches learning the value function Q via different algorithms, e.g., SARSA [74], MAXQ[34] or offline RL [97], using deep networks [162, 161] or simpler forms of value-function representations, e.g., look-up tables [74].

One challenge when using RL is that the robot needs to balance exploration and exploitation when interacting with the environment to collect rewards (see Chapter 2.3 for a more detailed discussion). To avoid catastrophic failures during exploration, safe RL methods emerged and have been applied to HRI. For example, related work explored how human experts who are actively involved at different stages of the training process could avert catastrophic actions. The concept of shielding [7] is one approach for safe RL and originally used linear temporal logic to implement high-level constraints for providing the agent with a set of safe actions. In HRI, the application of shields was used for discrete state and action spaces, like in cooking tasks for action modification through crowd sourcing [208], and to continuous state and action spaces for social navigation [124]. This thesis proposed to use shields to ensure socially appropriate behavior when learning backchanneling behavior (Paper E).

Through RL, HRI researchers develop novel forms of learning navigation or manipulation skills, for example through socially-guided machine learning [201] or preference learning [176, 92], focus on acquiring assistive social skills, e.g., to choose healthy nutritional drinks [170] or create reading [157] or letter learning curricula [9]. Other research focuses on using RL for learning social interaction skills, e.g., maintaining engagement [97, 33], or personalizing motivational strategies [74]. The remainder of this section will focus on methods for learning socially assistive and social interaction skills, which are the most relevant to this thesis.

RL has been used to personalize robot behavior to an interaction partner. For instance, Mitsunaga et al. [130] explored adaptive behavior to increase personal comfort based on human body signals. RL can also be used to adapt a robot's empathy [120], humor [216] and language [168] to comfort, provide entertainment, and improve learning outcomes [74]. Other work explored online RL to learn a policy for a humanoid robot to interact with bypassing strangers [162].

Most prior work in applying RL to HRI has been geared towards one-on-one interaction settings, except for work in robot navigation, which involves multiparty interactions by design [127]. Prior literature has also shown how to learn appropriate gaze behaviors when interacting in groups [212, 116]. This thesis builds upon those prior works using RL and IL in multiparty interactions but explores learning assistive social skills to improve interactions among adult second language learners and native speakers (Paper F).

Chapter 4

Computational approaches to interaction-shaping robotics

This chapter summarizes the contributed works that aim to develop autonomous social robots that are capable of interacting with multiple people and can dynamically adapt to human-human dynamics, positively shaping interactions between people in light of the two research questions:

RQ1 Can autonomous social robots capable of adapting to dynamic groups effectively shape interactions?

RQ2 How can we leverage machine learning methods to learn autonomous social robot behaviors that aim to shape interactions?

This chapter starts with providing a more in-depth definition and consideration of interaction-shaping robotics (Paper A) and with a formalization of the problem of interaction-shaping robot behavior as an MDP. The MDP captures people and their interactions as states and assumes actions that the robot can use to shape these interactions. This thesis then aims to demonstrate the effectiveness of autonomous, adapting, interaction-shaping robots using minimal perception and behavior heuristics. Section 4.3 first discusses the use of gaze to mediate participation imbalance in groups of adults (Paper B) before exploring the use of non-anthropomorphic robots to foster inclusion among children in a music game (Paper C). Lastly, this section explores what is needed for complex real-world interactions while co-designing interaction-shaping robots with teenagers. Section 4.4 then discusses how we could use online and offline learning to overcome the limitations of handcrafting behavior for learning social robot behaviors. The section first highlights how human-human datasets, in combination with methods from the safeRL community, could be used to allow for learning online, i.e., learning in interactions (Paper E). Lastly, this chapter discusses the use of offline learning methods, i.e., offline RL and IL, and explores the suitability of these methods for deploying socially appropriate robot behaviors that shape interactions among people (Paper F).

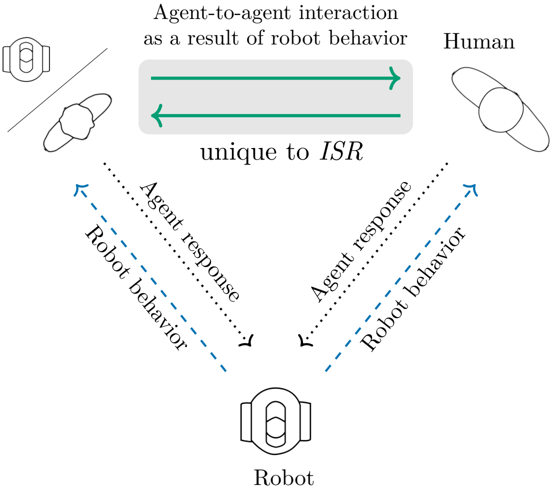
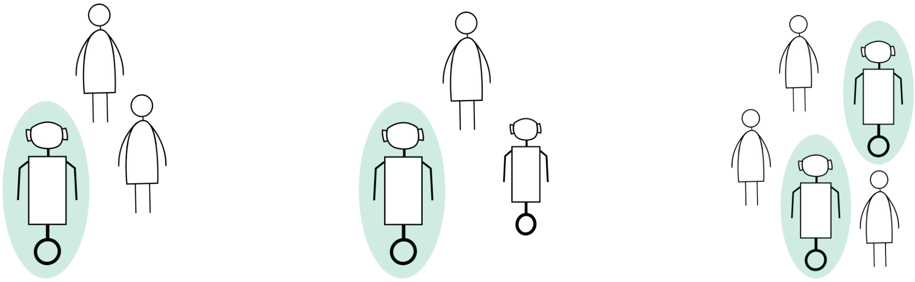


Figure 4.1: This schematic visualizes the interactions within a group of an interaction-shaping robot and two other agents. Blue-dashed arrows and black-dotted arrows show the interactions that are traditionally studied in the HRI community. Green solid arrows represent the interaction that is shaped by the robot, and we suggest is unique to ISR. Image source: Paper A [71]

4.1 Interaction-shaping robotics

At the core of this thesis is the notion of **interaction-shaping robotics** (ISR) which proposes a new subfield to HRI. ISR concerns the study of robots that shape interactions between other agents, e.g., between people, or between people and robots. Figure 4.1 visualizes the effects and phenomena traditionally studied in the HRI community through blue-dashed and black-dotted arrows. Green solid arrows capture the interaction between other agents that ISR is uniquely focusing on. For example, a robot might make a vulnerable statement by admitting its mistake in a game, which was found to ripple through the group and led to group members themselves making more vulnerable statements. The goal of this work and proposal of a new subfield to HRI was to provide an umbrella term that can align the terminology the community uses for works in this research area and to highlight the need for more works concerning ISR.

Based on prior related work capturing ISR scenarios, we identified five key factors that describe distinct methods used when interaction-shaping robots influence interactions between other agents: the role of the robot, the robot-shaping outcome, the form of robot influence, the type of robot communication, and the timeline of the robot's influence. As such, these factors serve to deepen our understanding of the different ways interaction-shaping robots can shape these interactions. An overview of the key factors is given in Table 3.1, and two key factors are discussed



(a) Structure I (b) Structure II (c) Structure III

Figure 4.2: Overview of the three interaction-shaping group compositions. Interaction-shaping robot(s) are marked with a green ellipse and shape the interactions among the remainder of agents. Image source: Paper A [71]

in Chapter 3.1.

Interaction-shaping robots might interact with other agents in a variety of different group compositions that might offer unique opportunities for these robots to have an impact. We propose three distinct compositions visualized in Figure 4.2. The first composition describes a group interaction as typically studied in this thesis. One robot shapes the interactions between two or more humans. In addition to the work in this thesis, this group composition has so far received the most attention in the community with studies capturing interaction-shaping effects among children [35], adults [195] and families [78]. Robots that shape interactions within a mixed group of humans and robots are studied in the second group composition. In this composition, a robot might, for example, share information about another robot's capabilities with the human, e.g., 'It cannot hear you but see you' which then clearly would affect how the robot and human interact with each other. The broadest scenario is captured in the third group composition in which multiple robots shape interaction among multiple others which could be groups of people or mixed groups of people and robots. Interesting applications for this scenario have studied how swarms of robots might help the sharing of opinions among people [5].

In addition to reviewing the literature and showing opportunities to study the impact of interaction-shaping, we highlight computational, methodological, and ethical challenges. Ethical challenges discuss the positive impact of functioning interactions and relationships in light of the robot's conduct, such as nudging and deception, and the risk that biased robot behavior might entail. A more in-depth discussion is also provided at the end of this thesis in Chapter 5. We further argue and discuss the need for computational advances in line with the contributions and discussions in this thesis. Specifically, we argue that we need to advance computational aspects to achieve more robust interaction-shaping behavior that can account for unexpected human behaviors and evolving groups. Methodological works that concern ISR scenarios have yet to develop comparable study setups and measures and contribute datasets for computational approaches.

In summary, this work highlights the potential impact of social robots supporting people and their interactions and which design elements and study setups might be interesting to explore further. However, we also discuss that we can only achieve this impact if we carefully consider the methodologies used to arrive at our scientific conclusions, the maturity of the technology allowing for these robots to be deployed in the wild, in the places they might provide impact and the ethical concerns that will need to be considered.

Paper A: Interaction-shaping Robotics: Robots That Influence Interactions between Other Agents

Sarah Gillet, Marynel Vázquez, Sean Andrist, Iolanda Leite, Sarah Sebo.
In ACM Transactions on Human-Robot Interaction, Vol. 13, No. 1, 2024

Abstract: Work in Human-Robot Interaction (HRI) has investigated interactions between one human and one robot as well as human-robot group interactions. Yet, the field lacks a clear definition and understanding of the influence a robot can exert on interactions between other group members (e.g., human-to-human). In this paper, we define Interaction-Shaping Robotics (ISR), a subfield of HRI that investigates robots that influence the behaviors and attitudes exchanged between two (or more) other agents. We highlight key factors of Interaction-Shaping Robots that include the role of the robot, the robot-shaping outcome, the form of robot influence, the type of robot communication, and the timeline of the robot's influence. We also describe three distinct structures of human-robot groups to highlight the potential of ISR in different group compositions and discuss targets for a robot's interaction-shaping behavior. Finally, we propose areas of opportunity and challenges for future research in ISR.
Contributions by the author: SG initiated and led the collaboration among SG, MV, SA, IL and SS, and developed the initial suggestion for the key factors in close collaboration with SS. The final key factors were developed collaboratively among all authors. All authors contributed to the paper writing, SG took the lead on the section on ethical challenges, the introduction and description of key factors and further contributed to all other sections of the paper. All other authors contributed actively writing parts of the paper and giving feedback for revisions.

4.2 Formal problem definition

We define the problem of shaping interactions among people as an MDP. We define the set of states $s \in S$ capturing the human-human interaction, a set of actions $a \in A$ that aim to shape this interaction. and the transition function $T_a(s, s')$

describes how state s transitions into state s' as a result of action a. However, in this thesis, we treat $T_a(s, s')$ as unknown. The reward $r(s, a)$ in the MDP captures how the interaction improves, e.g. more balance in participation, as a reaction to the robot executing action a in state s. The reward $r(s, a)$ shall be larger/smaller if the interaction of the group is closer/more distant from the desired interaction. The robot's overall goal is then to maximize the reward over an interaction. Note that we will use imitation learning in parts of this thesis which does not rely on the reward. The robot shall then use a policy $\pi(s) \mapsto a$ that indicates which action a to take in a given state s. When building on heuristics $\pi(s) \mapsto a$ is handcrafted. However, in a learning setting, the robot will optimize for $\pi(s) \mapsto a$ to either maximize the reward in an RL setting or to follow the demonstrations as closely as possible in an IL context.

As one example for shaping interactions, this thesis explores balancing participation using a notion of *participation unevenness*. *Participation unevenness* [199] describes the distance between ideal equal participation and the actual distribution of participation. Formally, unevenness is defined as:

$$\text{uneven}_g = \sum_{i \in [0,..,|P|)} |\text{sp}^i - \overline{\text{sp}}| \tag{4.1}$$

with sp^i representing the amount of time that participant i has spoken over the total amount of speech of all human interactants. The term $\overline{\text{sp}}$ in eq. (4.1) corresponds to the mean of the relative speech time of all humans. That is, $\overline{\text{sp}} = \frac{1}{2} \sum_{i \in [0,..,|P|)} \text{sp}^i$.

Based on the definition of unevenness, we can define the measure of balance for a group g as a normalized variant of unevenness:

$$\text{bal}_g = 1 - \text{uneven}_g / \max(\text{uneven}_g) \tag{4.2}$$

4.3 Autonomous interaction-shaping robots - Effects and necessary features

This thesis firstly explores if interaction-shaping robots that autonomously adapt to the group and their dynamics can effectively shape interactions (Papers B and C). In addition, this thesis leverages co-design methods and human demonstrations to shed light on the crucial elements for autonomy that interaction-shaping robots should have to be meaningful to the target user group – teenagers (Paper D).

Mediating participation imbalance among people with different skill levels

To demonstrate the effectiveness of autonomous interaction-shaping robots, we explore how a robot's gaze behavior autonomously adapting to the perceived participation behavior in the group is effective when mediating an imbalance in the group.

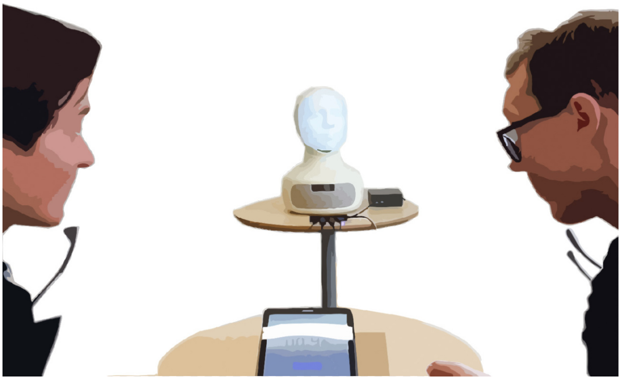


Figure 4.3: The interaction between the language learner (left), the native speaker (right) and the robot Furhat (middle). The two human group members describe words indicated on the tablet which the robot then has to guess.

We chose gaze as the shaping robot behavior because previous research found that 60-65% of human communication is non-verbal [31] which stresses the importance and potential for non-verbal behaviors such as gaze. In addition, pioneering work by Mutlu et al. [140, 139] demonstrated that participation roles in conversations (see 2 for a definition of different roles) can be shaped through human-like gaze behaviors and people take the turn following the gaze of the robot [187]. Different from prior work, we adapted the robot's gaze behavior moment-by-moment to the participation levels in the group.

The robot's adapting gaze behavior was studied with language learners and native speakers of Swedish in the "With other words" game as depicted in Figure 4.3. In the game, players described words – here, as a shared task for both the language learner and the native speaker – which the robot had to guess. We invited language learners and native speakers to create a natural imbalance in skill, which we expected to influence how much players participate through describing words. We increased the difficulty of describing the words over three difficulty levels throughout the game to ensure that the skill imbalance would occur despite varying skill among the language learners.

Given the task of describing words, we collected voice activation shares of the group members (collected automatically through individual close-talk microphones) to approximate participation behavior and measured the group dynamics through the unevenness in participation as given in Equation 4.1. In the user study, we further measured the number of turns taken, the personality and familiarity of participants, their language proficiency and willingness to communicate, as well as the social presence of the robot (see Section B.4.4 in Paper B for more details).

The user study then compared the adaptive and dynamic gaze behavior (DG) to a gaze behavior that was unaware of the group and group dynamics. The adaptive behavior algorithm led the robot to look more at the participant who talked less (see

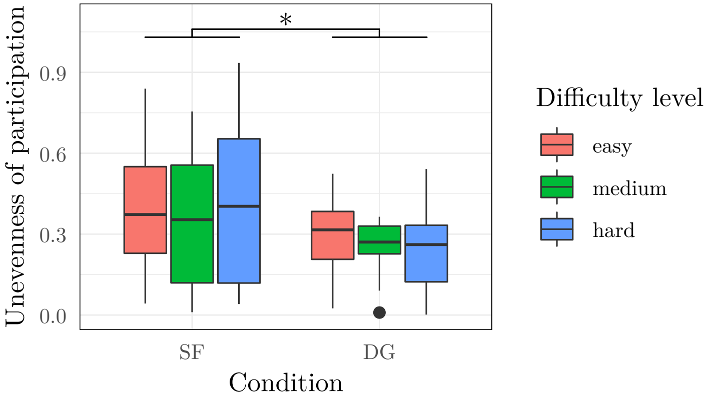


Figure 4.4: Unevenness in participation for the two gaze behaviors by word difficulty level (N=27 groups). Image adapted from: Paper B [69]

Section B.3 and Algorithm 1 in Paper B). The control gaze behavior, i.e., a speaker-focused gaze behavior (SF), combines looking at the current speaker with gaze aversion.

Figure 4.4 shows how the measure of unevenness in participation (see equation 4.1) resulted in lower values, meaning more even participation in the DG condition[1]. Additionally, we found that personality and language proficiency were a significant predictor of the amount of speech. There was no significant difference in the perception of social presence, suggesting that gaze behaviors were comparable and provided a fair comparison.

This work demonstrated the effectiveness of autonomous interaction-shaping robots that use minimal perception, i.e., voice activation shares, and dynamically adapting gaze behaviors. Even though the general amount of speaking was influenced by participants' proficiency and personality, the robot could effectively shape how participants interacted with each other, mediating the imbalance in the group. Thereby, our findings extend prior works that provided evidence for the power of gaze behaviors when influencing group interactions [140, 187]. Overall, we can conclude that gaze behaviors that acknowledge both group members according to their voice share should be used when robots aim to shape participation in skill-imbalanced groups.

[1] The original condition names were experimental (DG) and control (SF).

Paper B: Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels

Sarah Gillet*, Ronald Cumbal*, André Pereira,
José Lopes, Olov Engwall, Iolanda Leite.

In Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, 2021, New York, NY, USA. (HRI '21)

Abstract: Many small group activities, like working teams or study groups, have a high dependency on the skill of each group member. Differences in skill level among participants can affect not only the performance of a team, but also influence the social interaction of its members. In these circumstances, an active member could balance individual participation without exerting direct pressure on specific members by using indirect means of communication, such as gaze behaviors. Similarly, in this study, we evaluate whether a social robot can balance the level of participation in a language skill-dependent game, played by a native speaker and a second language learner. In a between-subjects study (N = 72), we compared an adaptive robot gaze behavior, that was targeted to increase the level of contribution of the least active player, with a non-adaptive gaze behavior. Our results imply that, while overall levels of speech participation were influenced predominantly by personal traits of the participants, the robot's adaptive gaze behavior could shape the interaction among participants which lead to more even participation during the game.

Contributions by the author: RC and SG ideated the project and were responsible for the conceptualization of the study, methodological design, code implementation, software development, participant recruitment and study execution. AP, JL, OE, and IL contributed by providing invaluable assistance and supervision across all phases of the research process. SG assumed the primary role in writing the paper, with RC collaborating in the process, and all authors contributing with important feedback and revisions.

Fostering inclusion among children

Factors that inhibit participation not only occur in crafted lab settings, as in the work discussed above, but can occur when children move to another country. One reason for the difficulty of participation might be a different language spoken in the new country. In addition, an increasingly polarized society and a growing impression of 'us vs. them' [164, 100] invite research on interaction-shaping robots that foster inclusion among children who newly arrived to a country and children already present.

For this purpose, we developed an interactive music-mixing game played by a group of three children - one newly arrived and two already present -, which was

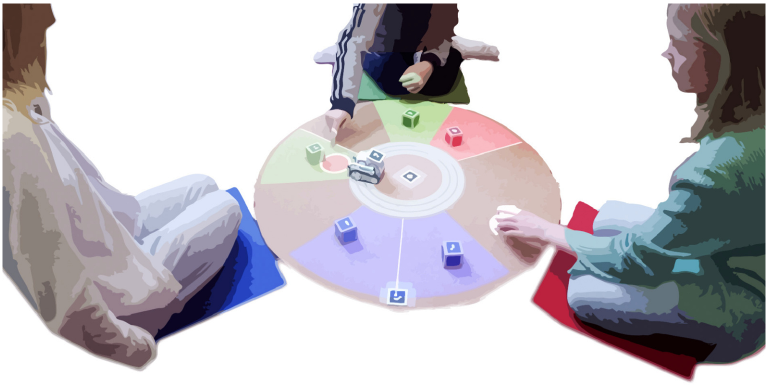


Figure 4.5: The robot Cozmo interacts with three children - two already present and one newly arrived - and measures their interaction dynamics through the movement of the tangible elements. It chooses the least exploring child as an addressee to encourage participation and exploration by prompting a replacement of the tangible elements.

mediated by the robot. The music-mixing game was designed as a language-free game to accommodate diverse language backgrounds. It involved placing tangible elements, i.e., cubes, on a round game board. The board was divided into nine wedges and each tangible element produced a different sound when placed in the different wedges. The goal of the game was to find the correct positions of the tangible elements indicated by rings lighting up in the middle of the board and a pleasant mix of the different sounds. Figure 4.5 shows the interaction between three children and the robot Cozmo, as well as, the game board.

The robot's goal was to achieve even participation and hence foster collaboration. Inspired by prior work [199], a directly-mediating behavior chose the child that had explored the least part of the board as the addressee of its action, i.e., the robot prompted the child to pick up and replace their tangible element. We compared this informed addressee choice with a random choice, i.e., the robot did not perceive the participation behavior and randomly chose which child to prompt. In addition to the directly mediating behavior, the robot displayed indirectly mediating behaviors by *following* previous cube movements and turning to the least exploring as an *encouraging* act.

In a between-subject study, 39 children, of which 13 had newly arrived in the country within the last two years, played the game with the Cozmo robot. The analysis of the children's interactions indicated that the robot could perceive the group's ingroup-outgroup dynamic and suggest respective displacements of tangible elements. Moreover, children in the mediation condition participated more evenly in the game even in a subsequent game round in which the robot was not present anymore. We further studied the effect of the robot's behavior beyond the game through a mini-dictator game [222, 77]. We found that children tended to be more prosocial, i.e., voluntarily give more stickers to other children, after interacting with the mediating robot.

This work showed that game-based approximations of group dynamics can be effective to perceive the group and choose actions that are sensible to overcome notions of ingroup-outgroup. Even though the robot's behavior was safe-guarded by a human controller, the autonomously chosen actions were never stopped by the controller. Therefore, this work further demonstrates the feasibility of developing autonomous interaction-shaping robots. This work extends prior work by using game-based approximations of group dynamics, demonstrating the feasibility of autonomy even for settings in which language and speech are not used or not indicative of the group dynamics.

Paper C: A social robot mediator to foster collaboration and inclusion among children

Sarah Gillet, Wouter van den Bos, Iolanda Leite.
In Proceedings of Robotics: Science and Systems, 2020, Corvalis, Oregon, USA. (RSS '20)

Abstract: Formation of subgroups and thereby the problem of intergroup bias is well-studied in psychology. Already from the age of five, children can show ingroup preferences. We developed a social robot mediator to explore how a robot could help overcome these intergroup biases, especially for children newly arrived to a country. By utilizing an online evaluation of collaboration levels, we allow the robot to perceive and act upon the current group dynamics. We investigated the effectiveness of the robot's mediating behavior in a between-subject study with 39 children, of whom 13 children had arrived in Sweden within the last 2 years. Results indicate that the robot could help the process of inclusion by mediating the activity. The robot succeeds in encouraging the newly arrived children to act more outgoing and in increasing collaboration among ingroup children. Further, children show a higher level of prosociality after interacting with the robot. In line with prior work, this study demonstrates the ability of social robotic technology to assist group processes.

Summary of Contributions: Summary of contributions.

Contributions by the author: SG and IL ideated the project. SG designed and implemented the game and robot behaviors and performed the evaluation. The study design was developed by SG in collaboration with WB and IL. SG wrote the paper with feedback from WB and IL.

Working with stakeholders to define robot behaviors and sensing

Researchers investigating robots that support small group interactions build robot behaviors informed by social psychology research [195, 193], by experts [131] or

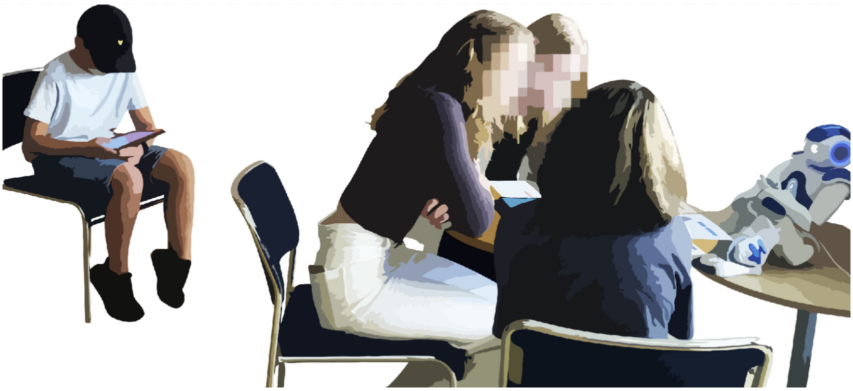


Figure 4.6: The interaction between the robot and three teenagers in a discussion-based task (right) and the teenager wizarding the robot through the actions co-designed by the group (left).

through co-design with stakeholders [156]. Papers B and C used handcrafted autonomous behavior policies according to best practices from the literature. Whereas these handcrafted behaviors were effective, other works have found that effects might differ from those intended [193, 105, 186], e.g., a robot designed to mediate a conflict instead called attention to this conflict [105]. Therefore, we also considered involving stakeholders in the development of interaction-shaping robots. In our case, we involved teenagers.

Involving teenagers allowed us to (a) understand their wishes for and views on the robot and (2) understand how they would like to interact with the robot. The goal of our co-design process on behaviors and sensing was to define the actions $a \in \mathcal{A}$ of the robot and the state space $s \in \mathcal{S}$ and collect demonstrations of $\pi : s \mapsto a$.

We invited 16 teenagers (8 boys, 8 girls, ages 12-15 with M=12.8 years old) divided into three groups to explore with us which social robot behaviors are needed and desirable for better group interactions in their peer groups. We did not prime them on what it meant to have a "better" group interaction but aimed to capture their spontaneous interpretations. We undertook different focus groups and design activities as visualized in Figure 4.7 and facilitated the design of a concrete action space.

As part of our robot design activities, we invited teenagers to use the designed action space in group discussion sessions. In these sessions, three teenagers worked on a discussion-based task such as “Which robot for the job?”[2]. One of the teenagers controlled the robot Nao (See Figure 4.6) and one observed the interaction to reflect on the robot’s action design as proposed by [22, 23].

[2] For this task, we provided job descriptions and robot pictures and asked them to discuss which robot would be the most suitable to be “hired” for this job.

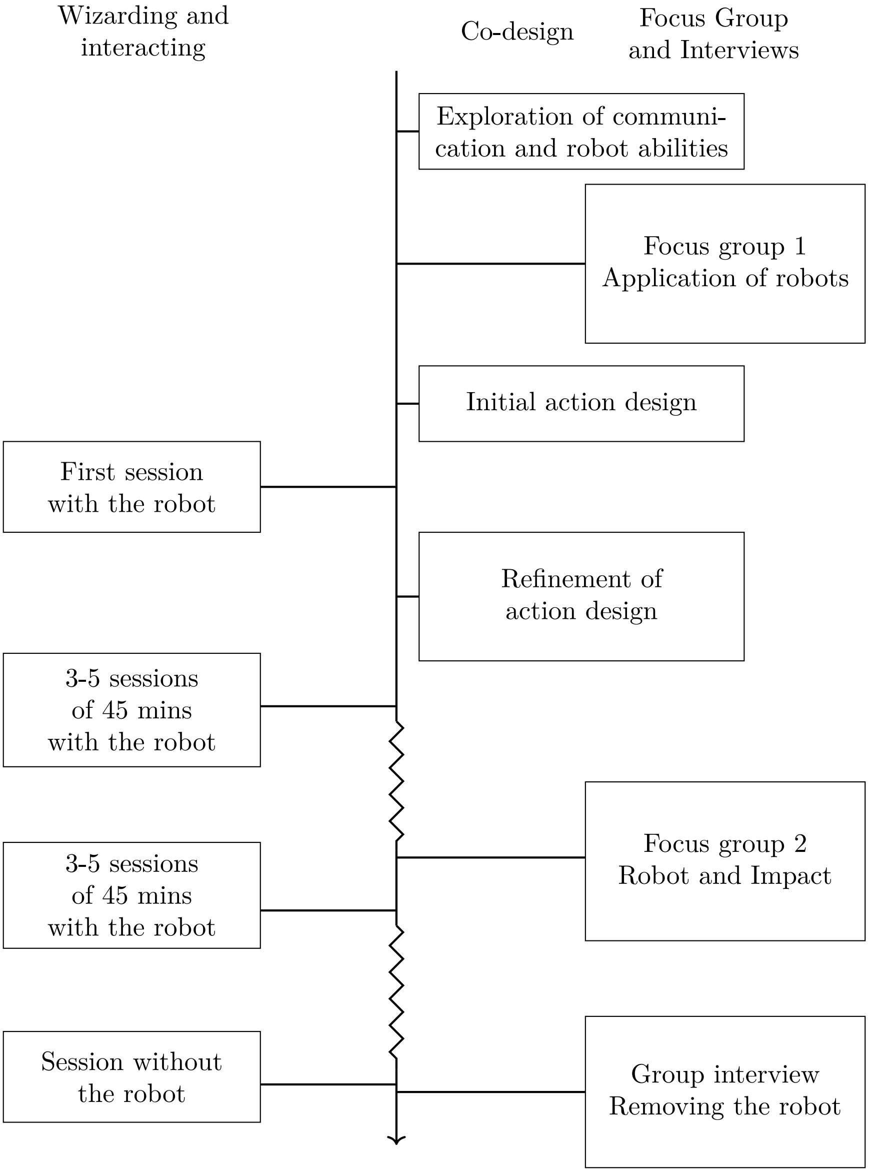

Wizarding and interacting
Co-design
Focus Group and Interviews
Exploration of communication and robot abilities
Focus group 1 Application of robots
Initial action design
First session with the robot
Refinement of action design
3-5 sessions of 45 mins with the robot
Focus group 2 Robot and Impact
3-5 sessions of 45 mins with the robot
Session without the robot
Group interview Removing the robot

Figure 4.7: Timeline of the co-design and interaction sessions for understanding teenagers' wishes and uses of a social robot that supports their group interactions.

Our analysis of the focus group discussion suggests that all three groups could imagine a robot being useful to break the ice and to support equal participation and turn taking, which was also reflected in the action space teenagers designed. In addition, teenagers imagined the robot to serve as a tool for improving efficiency and performance, e.g., by answering questions through internet searches.

The three different groups used the robot very differently during the sessions with the designed action space on the robot. The robot controlled in the group we called “Baymax” used the robot more like a teacher, asking the group to focus and give reminders to collaborate. On the other extreme, the “R2D2” group used the robot to make the activity more fun and entertaining. The third group “WALLE” could be placed in between with some fun elements but also frequent requests to focus. We conclude that the varying dynamics in a group, which also develop over time, need different robot behavior policies to support the interactions. In addition, the differences between behavior policies demonstrated by the teenagers would make handcrafting simple autonomy through heuristics tedious. Therefore, this work contributes to the understanding of how teenagers would like to use robots for their group interactions and suggests that learning-based approaches might be needed for the robot to adapt to the different and evolving needs of these groups.

Paper D: Ice-Breakers, Turn-Takers and Fun-Makers: Exploring Robots for Groups with Teenagers

Sarah Gillet*, Katie Winkle*, Giulia Belgiovine*, Iolanda Leite.
In Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication, 2022, Naples, Italy. (RO-MAN 2022)

Abstract: Successful, enjoyable group interactions are important in public and personal contexts, especially for teenagers whose peer groups are important for self-identity and self-esteem. Social robots seemingly have the potential to positively shape group interactions, but it seems difficult to effect such impact by designing robot behaviors solely based on related (human interaction) literature. In this article, we take a user-centered approach to explore how teenagers envisage a social robot “group assistant”. We engaged 16 teenagers in focus groups, interviews, and robot testing to capture their views and reflections about robots for groups. Over the course of a two-week summer school, participants co-designed the action space for such a robot and experienced working with/wizarding it for 10+ hours. This experience further altered and deepened their insights into using robots as group assistants. We report results regarding teenagers’ views on the applicability and use of a robot group assistant, how these expectations evolved throughout the study, and their repeat interactions with the robot. Our results indicate that each group moves on a spectrum of need for the robot, reflected in use of the robot more (or less) for ice-breaking, turn-taking, and fun-making as the situation demanded.

Contributions by the author: KW ideated the project. SG and KW collaboratively refined the project idea and developed the study details and protocol. SG took responsibility for the robotic system and KW for the focus group protocols and interviews. SG, KW and GB jointly worked on the co-design approach, ran the summerschool, analyzed the data and wrote the paper. IL adviced SG, KW, and GB and gave feedback on the paper.

4.4 Learning appropriate interaction-shaping behaviors online and offline

Our work with teenagers, previous results of unexpected effects when using handcrafted behaviors, and the complexity of group interactions that would require tedious handcrafting, encouraged the exploration of learning-based approaches to interaction-shaping robotics. We chose to explore two types of learning settings – online in interaction and offline from datasets. In the following, this thesis first presents an approach to allow for online RL while staying socially appropriate. Lastly, approaches to offline learning – IL and RL– are presented in light of their effectiveness when mediating participation imbalance.

Toward learning in interaction through shielding

RL is a prominent approach for learning behavior policies for robots, in general, [110] often with the help of humans in the loop [123]. In this work, we explore shielding to ensure social appropriateness during the exploration of an RL agent.

Exploration is a crucial element of RL that enables the RL agent to later exploit on the collected rewards from the environment [196]. Exploration is based on taking random actions, which for a social robot could lead to inappropriate behavior, i.e., nodding continuously or gazing hastily between people. Therefore, we propose to use shielding, an approach from the safe RL community, to avoid inappropriate behaviors during exploration, learning, and deployment.

We decided to focus on attentive listening in the form of backchanneling, i.e., providing feedback through nods and small utterances in one-on-one interactions. BC behavior has been found to be idiosyncratic [25], i.e., it is unique to an individual. At the same time, however, research showed that backchanneling typically happens in backchanneling opportunity points (BOP) [89] (see also Section 2.1) of which people choose roughly a third for their backchannels. We use this notion of BOPs for our shielding approach, i.e., the shield should only allow backchannels in BOPs.

We use a conversational dataset [14] to explore if a data-driven approach to shielding could generate a shield that limits a randomly exploring RL agent to backchannel only in appropriate moments in a video-based study as shown in Figure 4.8. We compare the data-driven approach to two baselines: one random exploration behavior without a shield and a statistically guided behavior, which overall

Figure 4.8: We explore if shielding could allow for online RL for robot listening behavior. In the video study, the person tells a small story to which the robot generates listening behavior through backchannels, i.e., nods and small utterances.

limited the number of backchannels. The audio stimulus was kept constant for all three conditions so that participants would evaluate the robot and its listening behavior solely due to different robot behaviors.

The study with 92 participants recruited through Prolific showed that the shielded exploration behavior was perceived as a significantly closer and more comforting listener, but only when compared to the random exploration behavior. Interestingly, we found that the statistically guided and our shielded approach provide an acceptable interaction quality and that backchanneling in ”wrong“ moments might be acceptable as long as it is not too frequent. In future work, we aim to explore if these results from a video study transfer to in-person interactions and how different ways of shielding support the online RL process.

Paper E: Shielding for socially appropriate robot listening behaviors

Sarah Gillet, Daniel Marta, Mohammed Akif, Iolanda Leite.
Submitted to the 33rd IEEE International Conference on Robot and Human Interactive Communication, 2024, Pasadena, USA. (RO-MAN 2024)

Abstract: A crucial part of traditional reinforcement learning (RL) is the initial exploration phase, in which trying available actions randomly is a critical element. As random behavior might be detrimental to the interaction, this work proposes a novel paradigm for learning social robot behavior–the use of shielding to ensure socially appropriate behavior during exploration and learning. We explore how a data-driven approach for shielding could be used to generate listening behavior. In a video-based user study (N=110), we compare shielded exploration to two

other exploration methods. We show that the shielded exploration is perceived as more appropriate and comfortable than a straightforward random approach. Based on our findings from the user study, we discuss the potential for future work using shielded and socially guided approaches for learning idiosyncratic social robot behaviors through RL.

Contributions by the author: SG ideated the project and developed the approach in collaboration with DM. MA performed intial explorations of this work as part of his master thesis which was supervised by SG and DM. SG implemented the system. SG and MA designed the user study and SG prepared and ran the user study. SG led the paper writing efforts. DM and IL contributed to the writing of the paper. IL supervised SG and DM, evaluated MA's thesis and provided valuable feedback.

Balancing participation with learned behavior policies

We used the problem definition as given in Section 4.2 to formalize balancing participation through gaze as an MDP to be solved through IL and RL [70]. The goal was to train adjusting and flexible gaze policies $\pi(s) \mapsto a$ based on states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$, overcoming the workload and limitations that come with handcrafting behaviors.

To train the two policies, π_{RL} and π_{IL}, we first defined the state space $\mathcal{S}$ and action space $\mathcal{A}$ to extract the following dataset:

Interaction Context: The dataset comprises interactions between a language learner and a native speaker of Swedish playing the *With Other Words* game (Paper B). The two humans have to describe the current game word which the robot then has to guess within a fixed time. The robot's only crucial task is to guess the word so that the conversation evolves among the human players. Each pair of people played a minimum of 20 words and each session lasted 15-20 minutes.

State space: We provide a 77-element vector that encodes the state of human participants (36 features × 2), the state of the robot (3 features), and high-level interaction information (2 features). All of these features are collected at 2 Hz. For each human participant, we use the audio signal captured through individual close-talk microphones to compute statistical quantities of mel-frequency cepstrum coefficients (MFCC), 4-dimensional prosody features, speech intensity, and pitch as well as the first derivative of these features over the last second. Further, we consider participation balance from the viewpoint of the participant and if the participant is currently talking as a feature. To capture the robot in the feature vector, we collect the current gaze target of the robot - the speaker, listener, or neither (looking away) for gaze aversion. To provide a sense of time and action frequency, we also capture the time past since the last robot gaze change and the frequency with which the robot changed gaze targets.

Action space: The actions $a_t \in A$ available in the dataset are discrete directions that the robot was gazing towards: *Look at speaker*, *Look at listener*, *Perform* *gaze aversion*, and *Do nothing.* Gaze aversion was only performed on the current speaker. The *Do nothing* action did not change the robot's state. In the dataset, the robot took action continuously, i.e., at 2Hz.

Robot behavior in the dataset: This dataset was collected from Paper B. Therefore, the data collected from eleven groups exemplifies the dynamic, experimental gaze condition which used actions *Look at speaker*, *Look at listener*, and *Do nothing.* The actions *Look at speaker*, *Perform gaze aversion*, and *Do nothing* are exemplified for the other fifteen groups in the speaker-focused, control condition.

Participants: 54 participants interacting in 27 groups form the dataset. The 12 groups experiencing the SL behavior were formed by 12 females, 11 males, and 1, rather not say, and the 15 groups interacting with the SF behavior of 12 females and 18 males with an average age of 31.79 years ($SD = 11.50$) and 32.03 years ($SD = 10.73$), respectively.

For training π_{RL}, we further acknowledged the importance of one conversational turn for balancing participation. During one conversational turn, the robot might attempt to shape the length of the turn but also encourage the other participant to take the turn. Therefore, the horizon for the RL problem was defined as one conversational turn, i.e., we learn a behavior policy π_{RL} that optimizes its behavior over one turn. We computed the reward to be proportional to the improvement over the balance of speech in the game (see equation 4.1) for training the policy π_{RL}, i.e., the reward compared the unevenness at the end of the previous turn to the unevenness at the end of this turn with the reward being positive if the unevenness was reduced and vice versa.

As we intend to learn policies from the dataset above, we used offline RL in the form of Double Deep Q-learning [85] to train π_{RL}. For imitation learning, i.e., behavioral cloning, we trained $\pi_{\mathrm{IL}} \to a$ and used the same dataset with the actions that the heuristics chose as the ground truth. To evaluate the trained policies, we used standard machine learning metrics for IL (Macro F1), and appropriate methods from the offline RL community to estimate the quality of the offline learned policy (Sequential Doubly Robust estimator). To compare the two policies prior to deployment, we used Weighted Importance Sampling (WIS), an additional method from the offline RL community, which is also applicable to π_{IL} when estimating the expected returns of the new policy given the dataset. Interestingly, the WIS expected π_{IL} to perform better than π_{RL} which, however, was not confirmed in the experiments.

Before deploying both policies in a user study, we analyzed the learned behavior policies π_{RL} and π_{IL} on a left-out part of the dataset to ensure appropriateness. We considered the frequency of action change, i.e., hastiness, the durations of actions, and the average time that each action was performed.

After verifying the general appropriateness of the behaviors, we compared the policy π_{RL} to the policy π_{IL} in a user study. None of the behavior policies improved over the heuristics regarding the balance in participation. Interestingly, participants in the RL condition took significantly more turns than participants in the original gaze aversion condition. In addition, we found that people spoke overall more in the IL condition than in the original experimental, speaker-listener condition. Our key takeaway from this work was that even though we did not find any differences in terms of balance, our results show promise for learning gaze behaviors for group interactions as interactions were not compromised.

Paper F: Learning Gaze Behaviors for Balancing Participation in Group Human-Robot Interactions

Sarah Gillet, Maria Teresa Parreira, Marynel Vázquez, Iolanda Leite.
In Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction, 2022, New York, NY, USA. (HRI '22)

Abstract: Robots can affect group dynamics. In particular, prior work has shown that robots that use hand-crafted gaze heuristics can influence human participation in group interactions. However, hand-crafting robot behaviors can be difficult and might have unexpected results in groups. Thus, this work explores learning robot gaze behaviors that balance human participation in conversational interactions. More specifically, we examine two techniques for learning a gaze policy from data: imitation learning (IL) and batch reinforcement learning (RL). First, we formulate the problem of learning a gaze policy as a sequential decision-making task focused on human turn-taking. Second, we experimentally show that IL can be used to combine strategies from hand-crafted gaze behaviors, and we formulate a novel reward function to achieve a similar result using batch RL. Finally, we conduct an offline evaluation of IL and RL policies and compare them via a user study (N=50). The results from the study show that the learned behavior policies did not compromise the interaction. Interestingly, the proposed reward for the RL formulation enabled the robot to encourage participants to take more turns during group human-robot interactions than one of the gaze heuristic behaviors from prior work. Also, the imitation learning policy led to more active participation from human participants than another prior heuristic behavior.

Contributions by the author: SG ideated the project, developed the problem formulation and offline RL approach. MTP was responsible for coding, testing and writing about the imitation learning gaze policy. SG and MTP ran the study and recruited participants jointly. SG led the paper writing with the help of MTB, MV and IL. MV and IL provided valuable advice throughout the project.

Chapter 5

Discussion

This thesis explores the generation of autonomous social robot behavior that can interact with and adapt to multiple people and positively shape their interactions. This chapter first discusses the thesis contributions in light of the research questions and then presents limitations and future work that concern the development of autonomous social interaction-shaping robots. This thesis then concludes with a critical discussion on the ethical challenges of robots shaping interactions highlighting the need for future work of computational but also empirical and psychological nature.

5.1 Generating autonomous interaction-shaping robot behavior

Robots that support and shape interactions between people have great potential. However, to allow for these robots to impact people and our societies, we need to develop autonomous systems that are capable of leveraging interaction-shaping effects in the real world. This thesis has explored several techniques that aim to take the first step toward real-world capable interaction-shaping robots.

Research Question 1, as posed in the introduction (see Chapter 1), aims to investigate autonomous interaction-shaping robots, their adaptation to dynamic groups, and the respective effects of these robots. The contributed works demonstrate that effective autonomy for interaction-shaping robots can be achieved by perceiving and adapting to the dynamics in group interactions. In Paper C, autonomy was achieved through perceiving game-related behaviors. These behaviors allowed the robot to extract *who* was playing/ exploring the least active and would benefit from support. We compared this informed supportive behavior to random, uninformed supportive behavior. Similarly, Paper B compared informed supportive behavior that was aware of all participants and dynamically addressed both of them through gaze to unaware behavior that just focused on the speaker. In both works, the informed, aware robot behavior led to effective improvements in the group interaction by fostering inclusion or mediating imbalances in participation.

Based on the findings of those two works, this thesis proposes the following hypothesis: **Autonomous robot behaviors for shaping interactions are effective if the robot has means to decide *who* should be supported with its behavior.** Paper B and Paper C provide evidence for this hypothesis in line with related work, for example, the MicBot turning toward the least active speaker to balance engagement in a discussion[199] or a decision facilitation robot prompting the least active for their opinion [181]. In Paper C, the robot used the awareness of the group dynamics to prompt the least exploring player to take an action in the game. Paper B explored how a robot choosing the least participating player in addition to generally considering both speakers can be effective. The latter highlights that, at times, addressing and supporting the group as a whole might be important. Interestingly, the need to support individuals and the whole group as such was also highlighted by stakeholders – the teenagers – when designing and using their robot for supportive group interactions as reported in Paper D. Future work will need to validate if perceiving *who* should be supported is sufficient for the effective shaping of human-human dynamics. Insights from related fields such as affective computing on techniques for perceiving the dominant participant [166] or approximate cohesion [182] might, for example, help to understand how perceiving the group to gain this insight on *who* to support can be achieved in various contexts.

Working with the teenagers on their group assistant further revealed that the ideal robot behavior might differ between groups and individuals. A close inspection of the usage pattern of the different groups showed that one group used the robot more as a teacher and another group used the robot to induce fun into their interactions. Handcrafting behaviors through a set of rules that meet the different needs of the different groups might be possible but tedious and labor intensive [152]. Therefore, this thesis further extends prior works' arguments for personalization and adaptation [99, 125, 17] to groups by hypothesizing that **Learning is needed to meet the needs of different and evolving groups and the individuals in these groups.**

Research Question 2 aims to explore how we could leverage machine learning methods with the goal of achieving more flexible behaviors, being able to adapt to more complex dynamics without extensive hand-crafting of behaviors. Paper E shows how we could allow for online learning by using shielding based on human-human data. The goal was to enable a robot to learn idiosyncratic robot behaviors by allowing for exploration while maintaining social appropriateness. We compared the shielded approach to a random almost continuously acting behavior and a statistically-guided behavior. The results indicate that maintaining social appropriateness either through shielding or by restricting the behavior statistically leads to significantly higher-rated listening behavior during the exploration phase of the RL agent. We therefore hypothesize that **It is important that the robot acts in a socially appropriate manner during learning and deployment.**

Paper F provides additional evidence for the importance of the social appropriateness of learned behaviors during deployment. This work proposes to use offline RL or IL on an HRI dataset to learn gaze policies that balance participation. An important step prior to deployment was to analyze the prominent behaviors and if they followed norms, i.e., looking more at the speaker than the listener, and how frequently the behaviors changed, i.e., how hastily the robot was changing its gaze target. The results indicate that learning is a promising approach because learned behaviors did not compromise the interaction. Interestingly, prior work found that acting inappropriately at times might contribute to the overall likability of the robot [128] but also highlights the risk that failures might negatively impact trust [203]. Therefore, future work will need to investigate if social appropriateness is key to the success of learned behaviors and how unintended and intended inappropriate behaviors might impact human-human dynamics.

Overall, this thesis provides initial evidence that different approaches might be promising for learning of interaction-shaping robot behaviors. Future work will need to further investigate these approaches and validate their promise in groups of varying sizes with varying needs. One way of improving the learning for shaping interactions in groups might be to model the interaction between participants explicitly through a graph. This modeling would then allow for using Graph Neural Networks that reason explicitly over the structure of the interaction. Future work might be able to then learn behaviors that can be effective in groups of varying sizes and, due to the structure, reasoning over individuals and their interactions, allow for capturing their nuanced needs.

Lastly, future work will need to explore how to evaluate and judge improved learned behaviors that might not have an effect in interaction. In the machine learning community, improvements of a few percent in accuracy over previous approaches define continuous scientific progress (see [159] for an example). Scientific progress in the HRI community is often defined through differences in participant behaviors and impressions of the robot, and a few percent of behavior improvement might, in fact, at most times, not be perceivable by participants, for example, when generating gestures for a listening agent [219]. In Paper E we saw that even though the robot was acting at theoretically inappropriate moments, it was acceptable as long as the frequency of acting was adhered to. While small improvements might not have measurable effects on interactions, step-wise progress could, over many iterations, achieve notable improvements. Future work will, therefore, need to consider when the improvements are worth the effort and cost to evaluate with users and how to present and publish the potentially necessary intermediate steps.

5.2 Limitations

The presented heuristics and learning methods could only be evaluated in the specific interaction contexts. Even though we attempted to design a more generic, real-world interaction scenario by bringing the robot to the school and playing the language game in a meeting room, our studies can be considered controlled lab studies. Our results are, therefore, limited to the chosen contexts. However, the presented approaches are general and might be applicable to other scenarios. In addition, the approaches were designed to mimic real-world interactions closely. Future work will need to further investigate the stated hypotheses and effects in real-world HRI.

We use computational measures of participation that might not reflect the actual contribution that participants made to the group. People might be able to contribute the same idea or progress to a discussion with much fewer or much more words and speech time. Therefore, the measure of participation as speech amount which we use in Paper B and Paper F might not actually capture their contribution. The choice for measuring contributions through speech amount was of a pure computational reason and the availability of this kind of data. Future work might explore how we could measure participation and group members' contributions in more appropriate ways, for example, by using Large Language Models.

5.3 Ethical challenges

Given the potential that positive human-human connections can have on general life satisfaction but also for equal opportunities, this thesis offers advancement that might bring us one step closer to actually using these robots as tools deployed in the real world. However, influencing people's behavior also raises ethical concerns. First of all, IEEE's guidelines on Ethically Aligned Design suggest that autonomous and intelligent systems should support human potential and ensure connections and relationships between humans [54]. The works that this thesis contributes aim to follow this suggestion by enhancing and promoting human-human connections. By leveraging the perception of a person's interaction dynamics, future work might further follow this guideline by, e.g., detecting if a person is isolated and encouraging connections between them and others. Even though promoting interactions between people might reduce the risk of isolation and the robot as an isolating factor [57], the risk of dependency on the robot cannot be fully eliminated. In Paper C, we therefore studied the impact of the robot in subsequent interactions without the robot. Future work should further focus on studying the robot's long-lasting effects going beyond the effects in the immediate interaction with the robot.

The robot behaviors explored in this thesis shape interactions through nonverbal behaviors, relying on sometimes intuitive reactions. As users might not be aware that a robot might influence their behaviors, interaction-shaping robots face the risk of deception. IEEE's ethics guidelines suggest that "In general, deception may be acceptable in an affective agent when it is used for the benefit of the person being deceived, not for the agent itself."[54, p. 175] In other words, deception in ISR might be ethically acceptable when the robot's shaping behavior benefits people as explored in this thesis. Nonetheless, future work should explore how reducing the ethical risk by being upfront about the robot's goal affects the robot's ability to shape interactions.

Chapter 6

Conclusion

The goal of this thesis is to advance the development of autonomous social robots that shape human-human interactions. First, this thesis proposes the terminology of interaction-shaping robotics and suggests that considering the current human-human dynamics is important when developing autonomous interaction-shaping robots. Especially, this thesis provides evidence that the perceived dynamics are important to decide *who* to address when fostering inclusion among children or supporting interactions in skill-imbalanced groups to balance participation. Further, intense work with teenagers demonstrated that when imagining and using the robot as a supporter for their interaction, the robot's role and behavior has to adapt to the varying and forming groups of teenagers which highlights that learning approaches might be needed to meet the needs of these groups. This thesis then explores the feasibility of machine learning methods that could overcome the limitations and labor of handcrafting heuristics. The contributed works show that learning behavior policies online might be feasible through the use of shielding to ensure socially appropriate exploration in the initial phases of online RL. Further, learning offline from human-robot interaction datasets might be feasible and not compromise the interaction when balancing participation in groups with different skill levels. Overall, this thesis contributes technical, user studies and theoretical work with the overall goal to develop robots that encourage and support people through the flourishing of their human-human interactions.

Bibliography

[1] Anna MH Abrams and Astrid M Rosenthal-von der Pütten. "I–C–E Framework: Concepts for Group Dynamics Research in Human-Robot Interaction". In: *International Journal of Social Robotics* (2020), pp. 1–17.

[2] Henny Admoni and Brian Scassellati. "Data-driven model of nonverbal behavior for socially assistive human-robot interactions". In: *Proceedings of the 16th international conference on multimodal interaction.* 2014, pp. 196–199.

[3] Neziha Akalin and Amy Loutfi. "Reinforcement Learning Approaches in Social Robotics". In: *Sensors* 21.4 (2021), p. 1292.

[4] Baris Akgun, Maya Cakmak, Jae Wook Yoo, and Andrea Lockerd Thomaz. "Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective". In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction.* 2012, pp. 391–398.

[5] Merihan Alhafnawi, Edmund R. Hunt, Severin Lemaignan, Paul O'Dowd, and Sabine Hauert. "MOSAIX: a Swarm of Robot Tiles for Social Human-Swarm Interaction". In: *2022 International Conference on Robotics and Automation (ICRA).* 2022, pp. 6882–6888.

[6] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. *The nature of prejudice.* Addison-wesley Reading, MA, 1954.

[7] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. "Safe reinforcement learning via shielding". In: *Proc. of the AAAI Conf. on Artificial Intelligence.* Vol. 32. 1. 2018.

[8] Patricia Alves-Oliveira, Pedro Sequeira, and Ana Paiva. "The role that an educational robot plays". In: *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016* August (2016), pp. 817–822.

[9] Aida Amirova, Nurziya Oralbayeva, Zhansaule Telisheva, Aida Zhanatkyzy, Aidar Shakerimov, Shamil Sarmonov, Arna Aimysheva, and Anara Sandygulova. "QWriter System for Robot-Assisted Alphabet Acquisition". In: *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* 2023, pp. 1227–1232.

[10] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. "A survey of robot learning from demonstration". In: *Robotics and autonomous systems* 57.5 (2009), pp. 469–483.

[11] Michael Argyle, Mark Cook, and Duncan Cramer. "Gaze and Mutual Gaze". In: *British Journal of Psychiatry* 165.6 (1994), pp. 848–850.

[12] Steven R Asher and Molly Stroud Weeks. "Loneliness and belongingness in the college years". In: *The handbook of solitude: Psychological perspectives on social isolation, social withdrawal, and being alone* (2013), pp. 283–301.

[13] Richard D Ashmore, Kay Deaux, and Tracy McLaughlin-Volpe. "An organizing framework for collective identity: articulation and significance of multidimensionality." In: *Psychological bulletin* 130.1 (2004), p. 80.

[14] Andrew J. Aubrey, David Marshall, Paul L. Rosin, Jason Vandeventer, Douglas W. Cunningham, and Christian Wallraven. "Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 2013, pp. 277–282.

[15] Christoph Bartneck and Jodi Forlizzi. "A design-centred framework for social human-robot interaction". In: *RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759).* IEEE. 2004, pp. 591–594.

[16] Roy F Baumeister and Mark R Leary. "The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation". In: *Psychological Bulletin* 117.3 (1995), pp. 497–529.

[17] Giulia Belgiovine, Jonas Gonzalez-Billandon, Giulio Sandini, Francesco Rea, and Alessandra Sciutti. "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. Barcelona, Spain: Association for Computing Machinery, 2022, pp. 139–145.

[18] Joseph Berger, David G Wagner, and Murray Webster Jr. "Expectation states theory: Growth, opportunities and challenges". In: *Advances in group processes* (2014), pp. 19–55.

[19] Aude Billard and Daniel Grollman. "Robot learning by demonstration". In: *Scholarpedia* 8.12 (2013), p. 3824.

[20] Frances R. Bilous and Robert M. Krauss. "Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads". In: *Language & Communication* 8.3 (1988). Special Issue Communicative Accomodation: Recent Developments, pp. 183–194.

[21] Chris Birmingham, Zijian Hu, Kartik Mahajan, Eli Reber, and Maja J Matarić. “Can I trust you? A user study of robot mediation of a support group”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 8019–8026.

[22] Elin A Björling and Emma Rose. “Participatory research principles in human-centered design: engaging teens in the co-design of a social robot”. In: *Multimodal Technologies and Interaction* 3 (2019).

[23] Elin A Björling, Kyle Thomas, Emma J Rose, and Maya Cakmak. “Exploring teens as robot operators, users and witnesses in the wild”. In: *Frontiers in Robotics and AI* 7 (2020), p. 5.

[24] Peter Blomsma, Gabriel Skantze, and Marc Swerts. “Backchannel Behavior Influences the Perceived Personality of Human and Artificial Communication Partners”. In: *Frontiers in Artificial Intelligence* 5 (2022).

[25] Peter Blomsma, Julija Vaitonytė, Gabriel Skantze, and Marc Swerts. “Backchannel behavior is idiosyncratic”. In: *Language and Cognition* (2024), pp. 1–24.

[26] Dan Bohus, Chit W Saw, and Eric Horvitz. “Directions robot: in-the-wild experiences and lessons learned”. In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 2014, pp. 637–644.

[27] Cynthia Breazeal. “Toward sociable robots”. In: *Robotics and autonomous systems* 42.3-4 (2003), pp. 167–175.

[28] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. “Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment”. In: *Journal of Human-Robot Interaction* 2.1 (2013), pp. 82–111.

[29] Marilynn B Brewer. “In-group bias in the minimal intergroup situation: A cognitive-motivational analysis.” In: *Psychological bulletin* 86.2 (1979), p. 307.

[30] Judee K Burgoon, Deborah A Coker, and Ray A Coker. “Communicative effects of gaze behavior: A test of two contrasting explanations”. In: *Human Communication Research* 12.4 (1986), pp. 495–524.

[31] Judee K. Burgoon, David B. Buller, Jerold L. Hale, and Mark A. de Turck. “Relational Messages Associated With Nonverbal Behaviors”. In: *Human Communication Research* 10.3 (1984), pp. 351–378.

[32] Donald T. Campbell. “Common fate, similarity, and other indices of the status of aggregates of persons as social entities”. In: *Behavioral Science* 3.1 (1958), pp. 14–25.

[33] Jeanie Chan and Goldie Nejat. “A learning-based control architecture for an assistive robot providing social engagement during cognitively stimulating activities”. In: *2011 IEEE International Conference on Robotics and Automation*. IEEE. 2011, pp. 3928–3933.

[34] Jeanie Chan and Goldie Nejat. “Social intelligence for a robot engaging people in cognitive training activities”. In: *International Journal of Advanced Robotic Systems* 9.4 (2012), p. 113.

[35] Vicky Charisi, Luis Merino, Marina Escobar, Fernando Caballero, R Gomez, and E Gómez. “The Effects of Robot Cognitive Reliability and Social Positioning on Child-Robot Team Dynamics”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA).* IEEE. 2021.

[36] Michael Jae-Yoon Chung and Maya Cakmak. ““How was Your Stay?”: Exploring the Use of Robots for Gathering Customer Feedback in the Hospitality Industry”. In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE. 2018, pp. 947–954.

[37] Herbert H Clark. *Using language.* Cambridge university press, 1996.

[38] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. “Multi-armed bandits with fairness constraints for distributing resources to human teammates”. In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.* 2020, pp. 299–308.

[39] Houston Claure, Seyun Kim, René F Kizilcec, and Malte Jung. “The social consequences of machine allocation behavior: Fairness, interpersonal perceptions and performance”. In: *Computers in Human Behavior* 146 (2023), p. 107628.

[40] Joe Connolly, Viola Mocz, Nicole Salomons, Joseph Valdez, Nathan Tsoi, Brian Scassellati, and Marynel Vázquez. “Prompting prosocial human interventions in response to robot mistreatment”. In: *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction.* 2020, pp. 211–220.

[41] Filipa Correia, Samuel Mascarenhas, Rui Prada, Francisco S. Melo, and Ana Paiva. “Group-based Emotions in Teams of Humans and Robots”. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI ’18* February (2018), pp. 261–269.

[42] Filipa Correia, Samuel F Mascarenhas, Samuel Gomes, Patrı́écia Arriaga, Iolanda Leite, Rui Prada, Francisco S Melo, and Ana Paiva. “Exploring prosociality in human-robot teams”. In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE. 2019, pp. 143–151.

[43] Richard S Crutchfield. “Conformity and character.” In: *American psychologist* 10.5 (1955), p. 191.

[44] Kenneth L Dion. “Group cohesion: From" field of forces" to multidimensional construct.” In: *Group Dynamics: Theory, research, and practice* 4.1 (2000), p. 7.

[45] Malcolm Doering, Dražen Brščić, and Takayuki Kanda. “Data-Driven Imitation Learning for a Shopkeeper Robot with Periodically Changing Product Information”. In: *J. Hum.-Robot Interact.* 10.4 (2021).

[46] Malcolm Doering, Dylan F. Glas, and Hiroshi Ishiguro. "Modeling Interaction Structure for Robot Imitation Learning of Human Social Behavior". In: *IEEE Transactions on Human-Machine Systems* 49.3 (2019), pp. 219–231.

[47] Paul Downing, Chris Dodds, and David Bray. "Why does the gaze of others direct visual attention?" In: *Visual Cognition* 11.1 (2004), pp. 71–79.

[48] Jon Driver IV, Greg Davis, Paola Ricciardelli, Polly Kidd, Emma Maxwell, and Simon Baron-Cohen. "Gaze perception triggers reflexive visuospatial orienting". In: *Visual cognition* 6.5 (1999), pp. 509–540.

[49] Erik Ekstedt and Gabriel Skantze. "Voice Activity Projection: Self-supervised Learning of Turn-taking Events". In: *INTERSPEECH 2022*. International Speech Communication Association. 2022, pp. 5190–5194.

[50] Olov Engwall, José Lopes, and Anna Åhlund. "Robot Interaction Styles for Conversation Practice in Second Language Learning". In: *International Journal of Social Robotics* (2020).

[51] Hadas Erel, Elior Carsenti, and Oren Zuckerman. "A Carryover Effect in HRI: Beyond Direct Social Effects in Human-Robot Interaction". In: *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '22. Sapporo, Hokkaido, Japan: IEEE Press, 2022, pp. 342–352.

[52] Hadas Erel, Yoav Cohen, Klil Shafrir, Sara Daniela Levy, Idan Dov Vidra, Tzachi Shem Tov, and Oren Zuckerman. "Excluded by Robots: Can Robot-Robot-Human Interaction Lead to Ostracism?" In: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '21. Boulder, CO, USA: Association for Computing Machinery, 2021, pp. 312–321.

[53] Hadas Erel, Denis Trayman, Chen Levy, Adi Manor, Mario Mikulincer, and Oren Zuckerman. "Enhancing Emotional Support: The Effect of a Robotic Object on Human–Human Support Quality". In: *International Journal of Social Robotics* (2021), pp. 1–20.

[54] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2*. IEEE, 2017.

[55] Vanessa Evers, Nuno Menezes, Luis Merino, Dariu Gavrila, Fernando Nabais, Maja Pantic, Paulo Alvito, and Daphne Karreman. "The development and real-world deployment of frog, the fun robotic outdoor guide". In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 2014, pp. 100–100.

[56] Friederike Eyssel and Dieta Kuchenbrandt. "Social categorization of social robots: Anthropomorphism as a function of robot group membership". In: *British Journal of Social Psychology* 51.4 (2012), pp. 724–731.

[57] David Feil-Seifer and Maja J. Matarić. “Defining socially assistive robotics”. In: *Proceedings of the 2005 IEEE 9th International Conference on Rehabilitation Robotics* 2005.Ci (2005), pp. 465–468.

[58] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. “A survey of socially interactive robots”. In: *Robotics and autonomous systems* 42.3-4 (2003), pp. 143–166.

[59] D.R. Forsyth. *Group Dynamics.* Cengage Learning, 2018.

[60] Marlena R Fraune, Selma Šabanović, and Takayuki Kanda. “Human group presence, group characteristics, and group norms affect human-robot interaction in naturalistic settings”. In: *Frontiers in Robotics and AI* 6 (2019), p. 48.

[61] Marlena R. Fraune, Selma Sabanovic, and Eliot R. Smith. “Teammates first: Favoring ingroup robots over outgroup humans”. In: *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication* 2017-Janua.August (2017), pp. 1432–1437.

[62] John RP French, Bertram Raven, et al. “The bases of social power”. In: *Studies in social power* 150 (1959), p. 167.

[63] Chris Kelland Friesen, Jelena Ristic, and Alan Kingstone. “Attentional effects of counterpredictive gaze and arrow cues.” In: *Journal of Experimental Psychology: Human Perception and Performance* 30.2 (2004), p. 319.

[64] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. “Gaze cueing of attention: visual attention, social cognition, and individual differences.” In: *Psychological bulletin* 133.4 (2007), p. 694.

[65] Scott Fujimoto, David Meger, and Doina Precup. “Off-Policy Deep Reinforcement Learning without Exploration”. In: *Proceedings of the 36th International Conference on Machine Learning.* Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 2052–2062.

[66] Chris Fullwood and Gwyneth Doherty-Sneddon. “Effect of gazing at the camera during a video link on recall”. In: *Applied Ergonomics* 37.2 (2006), pp. 167–175.

[67] Javier Garcıa and Fernando Fernández. “A comprehensive survey on safe reinforcement learning”. In: *Journal of Machine Learning Research* 16.1 (2015), pp. 1437–1480.

[68] Daniel Gatica-Perez. “Automatic nonverbal analysis of social interaction in small groups: A review”. In: *Image and Vision Computing* 27.12 (2009). Visual and multimodal analysis of human spontaneous behaviour: pp. 1775–1787.

[69] Sarah Gillet, Ronald Cumbal, André Pereira, José Lopes, Olov Engwall, and Iolanda Leite. "Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels". In: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* New York, NY, USA: ACM, 2021. ISBN: 9781450382892.

[70] Sarah Gillet, Maria Teresa Parreira, Marynel Vázquez, and Iolanda Leite. "Learning Gaze Behaviors for Balancing Participation in Group Human-Robot Interactions". In: *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction.* HRI '22. Sapporo, Hokkaido, Japan: IEEE Press, 2022, pp. 265–274.

[71] Sarah Gillet, Marynel Vázquez, Sean Andrist, Iolanda Leite, and Sarah Sebo. "Interaction-Shaping Robotics: Robots That Influence Interactions between Other Agents". In: *J. Hum.-Robot Interact.* 13.1 (2024).

[72] Arthur M Glenberg, Jennifer L Schroeder, and David A Robertson. "Averting the gaze disengages the environment and facilitates remembering". In: *Memory & cognition* 26 (1998), pp. 651–658.

[73] ERVING GOFFMAN. In: *Semiotica* 25.1-2 (1979), pp. 1–30.

[74] Goren Gordon, Samuel Spaulding, Jacqueline Korywestlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. "Affective personalization of a social robot tutor for children's second language skills". In: *30th AAAI Conference on Artificial Intelligence, AAAI 2016* 2011 (2016), pp. 3951–3957.

[75] Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, R. J. van der Werf, and Louis-Philippe Morency. "Virtual Rapport". In: *Intelligent Virtual Agents.* Ed. by Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 14–27.

[76] Victoria Groom and Clifford Nass. "Can robots be teammates?: Benchmarks in human–robot teams". In: *Interaction Studies* 8.3 (2007), pp. 483–500.

[77] Berna Güroglu, Wouter van den Bos, and Eveline A. Crone. "Sharing and giving across adolescence: An experimental study examining the development of prosocial behavior". In: *Frontiers in Psychology* 5.APR (2014), pp. 1–13.

[78] Omer Gvirsman and Goren Gordon. "Effect of Social Robot's Role and Behavior on Parent-Toddler Interaction". In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction.* HRI '24. <conf-loc>, <city>Boulder</city>, <state>CO</state>, <country>USA</country>, </conf-loc>: Association for Computing Machinery, 2024, pp. 222–230.

[79] Omer Gvirsman, Yaacov Koren, Tal Norman, and Goren Gordon. "Patricc: A platform for triadic interaction with changeable characters". In: *ACM/IEEE International Conference on Human-Robot Interaction* (2020), pp. 399–407.

[80] Victor H. Yngve. “On getting a word in edgewise”. In: Chicago Linguistic Society, 1970, pp. 567–577.

[81] J Richard Hackman. *Leading teams: Setting the stage for great performances.* Harvard Business Press, 2002.

[82] William F Hanks. *Language and communicative practices.* Routledge, 2018.

[83] Robert E Harlow and Nancy Cantor. “Still participating after all these years: a study of life task participation in later life.” In: *Journal of personality and social psychology* 71.6 (1996), p. 1235.

[84] Michelle A Harris and Ulrich Orth. “The link between self-esteem and social relationships: A meta-analysis of longitudinal studies.” In: *Journal of personality and social psychology* 119.6 (2020), p. 1459.

[85] Hado van Hasselt, Arthur Guez, and David Silver. “Deep Reinforcement Learning with Double Q-Learning”. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.* AAAI’16. Phoenix, Arizona: AAAI Press, 2016, pp. 2094–2100.

[86] Nick Hawes, Christopher Burbridge, Ferdian Jovan, Lars Kunze, Bruno Lacerda, Lenka Mudrova, Jay Young, Jeremy Wyatt, Denise Hebesberger, Tobias Kortner, et al. “The strands project: Long-term autonomy in everyday environments”. In: *IEEE Robotics & Automation Magazine* 24.3 (2017), pp. 146–156.

[87] Kotaro Hayashi, Daisuke Sakamoto, Takayuki Kanda, Masahiro Shiomi, Satoshi Koizumi, Hiroshi Ishiguro, Tsukasa Ogasawara, and Norihiro Hagita. “Humanoid robots as a passive-social medium-a field experiment at a train station”. In: *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE. 2007, pp. 137–144.

[88] Frank Hegel, Claudia Muhl, Britta Wrede, Martina Hielscher-Fastabend, and Gerhard Sagerer. “Understanding social robots”. In: *2009 Second International Conferences on Advances in Computer-Human Interactions.* IEEE. 2009, pp. 169–174.

[89] Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. “Backchannel relevance spaces”. In: *Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012.* Peter Lang Publishing Group. 2013, pp. 137–146.

[90] Guy Hoffman, Oren Zuckerman, Gilad Hirschberger, Michal Luria, and Tal Shani Sherman. “Design and evaluation of a peripheral robotic conversation companion”. In: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction.* 2015, pp. 3–10.

[91] Jess Hohenstein, Rene F Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F Jung. “Artificial intelligence in communication impacts language and social relationships”. In: *Scientific Reports* 13.1 (2023), p. 5487.

[92] Simon Holk, Daniel Marta, and Iolanda Leite. “PREDILECT: Preferences Delineated with Zero-Shot Language-based Reasoning in Reinforcement Learning”. In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction.* 2024, pp. 259–268.

[93] Julianne Holt-Lunstad. “Why social relationships are important for physical health: A systems approach to understanding and modifying risk and protection”. In: *Annual review of psychology* 69 (2018), pp. 437–458.

[94] Robert J House, Paul J Hanges, Mansour Javidan, Peter W Dorfman, and Vipin Gupta. *Culture, leadership, and organizations: The GLOBE study of 62 societies.* Sage publications, 2004.

[95] Lixing Huang and Jonathan Gratch. “Explaining the Variability of Human Nonverbal Behaviors in Face-to-Face Interaction”. In: *Intelligent Virtual Agents.* Ed. by Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 275–284.

[96] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. “Virtual Rapport 2.0”. In: *Intelligent Virtual Agents.* Ed. by Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 68–79.

[97] Nusrah Hussain, Engin Erzin, T. Metin Sezgin, and Yücel Yemez. “Training Socially Engaging Robots: Modeling Backchannel Behaviors with Batch Reinforcement Learning”. In: *IEEE Transactions on Affective Computing* 13.4 (2022), pp. 1840–1853.

[98] Tariq Iqbal and Laurel D. Riek. “Coordination dynamics in multihuman multirobot teams”. In: *IEEE Robotics and Automation Letters* 2.3 (2017), pp. 1712–1717.

[99] Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Dylan F. Glas, Iolanda Leite, and Kheng Lee Koay. “Personalization in Long-Term Human-Robot Interaction”. In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* 2019, pp. 685–686.

[100] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. “Affect, Not Ideology: A Social Identity Perspective on Polarization”. In: *Public Opinion Quarterly* 76.3 (2012), pp. 405–431.

[101] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. “Exploring Semi-Supervised Learning for Predicting Listener Backchannels”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 2021, pp. 1–12.

[102] Jeroen Janssen and Daniel Bodemer. “Coordinated computer-supported collaborative learning: Awareness and awareness tools”. In: *Educational psychologist* 48.1 (2013), pp. 40–55.

[103] Malte F Jung, Selma Šabanović, Friederike Eyssel, and Marlena Fraune. “Robots in groups and teams”. In: *Companion of the 2017 ACM conference on computer supported cooperative work and social computing.* 2017, pp. 401–407.

[104] Malte F. Jung, Dominic Difranzo, Solace Shen, Brett Stoll, Houston Claure, and Austin Lawrence. “Robot-Assisted Tower Construction—A Method to Study the Impact of a Robot’s Allocation Behavior on Interpersonal Dynamics and Collaboration in Groups”. In: *J. Hum.-Robot Interact.* 10.1 (2020).

[105] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. “Using Robots to Moderate Team Conflict: The Case of Repairing Violations”. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction.* Portland, Oregon, USA: Association for Computing Machinery, 2015, pp. 229–236.

[106] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. “Interactive robots as social partners and peer tutors for children: A field trial”. In: *Human–Computer Interaction* 19.1-2 (2004), pp. 61–84.

[107] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. “A two-month field trial in an elementary school for long-term human–robot interaction”. In: *IEEE Transactions on robotics* 23.5 (2007), pp. 962–971.

[108] Adam Kendon. “Some functions of gaze-direction in social interaction”. In: *Acta psychologica* 26 (1967), pp. 22–63.

[109] Chris L Kleinke. “Gaze and eye contact: a research review.” In: *Psychological bulletin* 100.1 (1986), p. 78.

[110] Jens Kober, J Andrew Bagnell, and Jan Peters. “Reinforcement learning in robotics: A survey”. In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1238–1274.

[111] Dieta Kuchenbrandt, Friederike Eyssel, Simon Bobinger, and Maria Neufeld. “Minimal Group - Maximal Effect? Evaluation and Anthropomorphization of the Humanoid Robot NAO”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* Vol. 7072 LNAI. Coe 277. 2011, pp. 104–113.

[112] Dieta Kuchenbrandt, Friederike Eyssel, Simon Bobinger, and Maria Neufeld. “When a Robot’s Group Membership Matters”. In: *International Journal of Social Robotics* 5.3 (2013), pp. 409–417.

[113] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. “Gesticulator: A framework for semantically-aware speech-driven gesture generation”. In: *Proceedings of the 2020 International Conference on Multimodal Interaction.* ICMI ’20. Virtual Event, Netherlands: Association for Computing Machinery, 2020, pp. 242–250.

[114] Sascha Lange, Thomas Gabel, and Martin Riedmiller. “Batch Reinforcement Learning”. In: 2012, pp. 45–73.

[115] Stephen RH Langton and Vicki Bruce. “Reflexive visual orienting in response to the social attention of others”. In: *Visual cognition* 6.5 (1999), pp. 541–567.

[116] Stéphane Lathuilière, Benoit Massé, Pablo Mesejo, and Radu Horaud. “Neural network based reinforcement learning for audio–visual gaze control in human–robot interaction”. In: *Pattern Recognition Letters* 118 (2019), pp. 61–71.

[117] Colin Wayne Leach, Martijn Van Zomeren, Sven Zebel, Michael LW Vliek, Sjoerd F Pennekamp, Bertjan Doosje, Jaap W Ouwerkerk, and Russell Spears. “Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification.” In: *Journal of personality and social psychology* 95.1 (2008), p. 144.

[118] Iolanda Leite, Marissa McCoy, Monika Lohani, Daniel Ullman, Nicole Salomons, Charlene Stokes, Susan Rivers, and Brian Scassellati. “Emotional storytelling in the classroom: Individual versus group interaction between children and robots”. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 2015, pp. 75–82.

[119] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. “Comparing models of disengagement in individual and group interactions”. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 2015, pp. 99–105.

[120] Iolanda Leite, André Pereira, Ginevra Castellano, Samuel Mascarenhas, Carlos Martinho, and Ana Paiva. “Modelling Empathy in Social Robotic Companions”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7138 LNCS. July. 2012, pp. 135–147.

[121] Iolanda Leite, André Pereira, Ginevra Castellano, Samuel Mascarenhas, Carlos Martinho, and Ana Paiva. “Modelling empathy in social robotic companions”. In: *Advances in User Modeling: UMAP 2011 Workshops, Girona, Spain, July 11-15, 2011, Revised Selected Papers 19*. Springer. 2012, pp. 135–147.

[122] Leigh Levinson, Omer Gvirsman, Iris Melamed Gorodesky, Almogit Perez, Einat Gonen, and Goren Gordon. “Learning in Summer Camp with Social Robots: A Morphological Study: Studying Dynamics Using Social Robots”. In: *International Journal of Social Robotics* (2020).

[123] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. “A Review on Interactive Reinforcement Learning From Human Social Feedback”. In: *IEEE Access* 8 (2020), pp. 120757–120765.

[124] Daniel Marta, Christian Pek, Gaspar I Melsión, Jana Tumova, and Iolanda Leite. “Human-feedback shield synthesis for perceived safety in deep reinforcement learning”. In: *IEEE Robotics and Automation Letters* 7.1 (2021), pp. 406–413.

[125] Gonçalo S Martins, Luiés Santos, and Jorge Dias. “User-adaptive interaction in social robots: A survey focusing on non-physical interaction”. In: *International Journal of Social Robotics* 11 (2019), pp. 185–205.

[126] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, and Tetsunori Kobayashi. “Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant”. In: *Computer Speech and Language* 33.1 (2015), pp. 1–24.

[127] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. “Core Challenges of Social Robot Navigation: A Survey”. In: *arXiv preprint arXiv:2103.05668* (2021).

[128] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. “To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot”. In: *Frontiers in Robotics and AI* 4 (2017).

[129] Nicole Mirnig, Astrid Weiss, Gabriel Skantze, Samer Al Moubayed, Joakim Gustafson, Jonas Beskow, Björn Granström, and Manfred Tscheligi. “Face-to-face with a robot: What do we actually talk about?” In: *International Journal of Humanoid Robotics* 10.01 (2013), p. 1350011.

[130] Noriaki Mitsunaga, Christian Smith, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. “Robot Behavior Adaptation for Human-Robot Interaction based on Policy Gradient Reinforcement Learning”. In: *Journal of the Robotics Society of Japan* 24.7 (2006), pp. 820–829.

[131] Elinor Mizrahi, Noa Danzig, and Goren Gordon. “vRobotator: A Virtual Robot Facilitator of Small Group Discussions for K-12”. In: *Proc. ACM Hum.-Comput. Interact.* 6.CSCW2 (2022).

[132] Serge Moscovici. “Toward a theory of conversion behavior”. In: *Advances in experimental social psychology.* Vol. 13. Elsevier, 1980, pp. 209–239.

[133] Helen Mott and Helen Petrie. “Workplace interactions: Women’s linguistic behavior”. In: *Journal of Language and Social Psychology* 14.3 (1995), pp. 324–336.

[134] Carl L Mueller and Bradley Hayes. “Safe and Robust Robot Learning from Demonstration through Conceptual Constraints”. In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.* 2020, pp. 588–590.

[135] Anthony Mulac, Karen T Erlandson, W Jeffrey Farrar, Jennifer S Hallett, Jennifer L Molloy, and Margaret E Prescott. ““Uh-huh. What’s that all about?” Differing interpretations of conversational backchannels and questions as sources of miscommunication across gender boundaries”. In: *Communication Research* 25.6 (1998), pp. 641–668.

[136] Michael Murray, Nick Walker, Amal Nanavati, Patricia Alves-Oliveira, Nikita Filippov, Allison Sauppe, Bilge Mutlu, and Maya Cakmak. “Learning Backchanneling Behaviors for a Social Robot via Data Augmentation from Human-Human Conversations”. In: *Proceedings of the 5th Conference on Robot Learning.* Ed. by Aleksandra Faust, David Hsu, and Gerhard Neumann. Vol. 164. Proceedings of Machine Learning Research. PMLR, 2022, pp. 513–525.

[137] Saida Mussakhojayeva, Nazerke Kalidolda, and Anara Sandygulova. “Adaptive strategies for multi-party interactions with robots in public spaces”. In: *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9.* Springer. 2017, pp. 749–758.

[138] Bilge Mutlu and Jodi Forlizzi. “Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction”. In: *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE. 2008, pp. 287–294.

[139] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. “Conversational gaze mechanisms for humanlike robots”. In: *ACM Transactions on Interactive Intelligent Systems* 1.2 (2012), pp. 1–33.

[140] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. “Footing in human-robot conversations: how robots might shape participant roles using gaze cues”. In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction.* ACM. 2009, pp. 61–68.

[141] Isabel Neto, Filipa Correia, Filipa Rocha, Patricia Piedade, Ana Paiva, and Hugo Nicolau. “The Robot Made Us Hear Each Other: Fostering Inclusive Conversations among Mixed-Visual Ability Children”. In: *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction.* HRI ’23. tockholm, Sweden: Association for Computing Machinery, 2023, pp. 13–23.

[142] Isabel Neto, Yuhan Hu, Filipa Correia, Filipa Rocha, João Nogueira, Katharina Buckmayer, Guy Hoffman, Hugo Nicolau, and Ana Paiva. “" I’m Not Touching You. It’s The Robot!": Inclusion Through A Touch-Based Robot Among Mixed-Visual Ability Children”. In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction.* 2024, pp. 511–521.

[143] Isabel Neto, Wafa Johal, Marta Couto, Hugo Nicolau, Ana Paiva, and Arzu Guneysu. “Using tabletop robots to promote inclusive classroom experiences”. In: *Proceedings of the Interaction Design and Children Conference*. IDC ’20. London, United Kingdom: Association for Computing Machinery, 2020, pp. 281–292.

[144] Illah R Nourbakhsh, Clayton Kunz, and Thomas Willeke. “The mobot museum robot installations: A five year experiment”. In: *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*. Vol. 4. IEEE. 2003, pp. 3636–3641.

[145] David G Novick, Brian Hansen, and Karen Ward. “Coordinating turn-taking with gaze”. In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP’96*. Vol. 3. IEEE. 1996, pp. 1888–1891.

[146] Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, Joseph Mendelson, Jonas Beskow, and Joakim Gustafson. “Crowd-Sourced Design of Artificial Attentive Listeners.” In: 2017.

[147] Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, Kenneth Funes Mora, Jean-Marc Odobez, and Joakim Gustafson. “Towards an Engagement-Aware Attentive Artificial Listener for Multi-Party Interactions”. In: *Frontiers in Robotics and AI* 8 (2021).

[148] Catharine Oertel and Giampiero Salvi. “A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue”. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. 2013, pp. 99–106.

[149] Raquel Oliveira, Patrícia Arriaga, Filipa Correia, and Ana Paiva. “The Stereotype Content Model Applied to Human-Robot Interactions in Groups”. In: *ACM/IEEE International Conference on Human-Robot Interaction* 2019-March.March (2019), pp. 123–132.

[150] Raquel Oliveira, Patrıécia Arriaga, Patrıécia Alves-Oliveira, Filipa Correia, Sofia Petisca, and Ana Paiva. “Friends or foes? Socioemotional support and gaze behaviors in mixed groups of humans and robots”. In: (2018), pp. 279–288.

[151] Thomas Olsson, Pradthana Jarusriboonchai, Paweł Woźniak, Susanna Paasovaara, Kaisa Väänänen, and Andrés Lucero. “Technologies for enhancing collocated social interaction: review of design solutions and approaches”. In: *Computer Supported Cooperative Work (CSCW)* 29 (2020), pp. 29–83.

[152] Nurziya Oralbayeva, Amir Aly, Anara Sandygulova, and Tony Belpaeme. “Data-Driven Communicative Behaviour Generation: A Survey”. In: *ACM Transactions on Human-Robot Interaction* 13.1 (2024), pp. 1–39.

[153] Nurziya Oralbayeva, Amir Aly, Anara Sandygulova, and Tony Belpaeme. “Data-driven Communicative Behaviour Generation: A Survey”. In: *J. Hum.-Robot Interact.* 13.1 (2024).

[154] Jeff Orkin and Deb Roy. “Automatic learning and generation of social behavior from collective human gameplay”. In: *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1.* 2009, pp. 385–392.

[155] Jeff Orkin and Deb Roy. “The restaurant game: Learning social behavior and language from thousands of players online.” In: *J. Game Dev.* 3.1 (2008), pp. 39–60.

[156] Anastasia K. Ostrowski, Daniella DiPaola, Erin Partridge, Hae Won Park, and Cynthia Breazeal. “Older Adults Living With Social Robots: Promoting Social Connectedness in Long-Term Communities”. In: *IEEE Robotics & Automation Magazine* 26.2 (2019), pp. 59–70.

[157] Hae Won Park, Ishaan Grover, Samuel Spaulding, Louis Gomez, and Cynthia Breazeal. “A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education”. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 33. 01. 2019, pp. 687–694.

[158] Maria Teresa Parreira, Sarah Gillet, and Iolanda Leite. “Robot Duck Debugging: Can Attentive Listening Improve Problem Solving?” In: *Proceedings of the 25th International Conference on Multimodal Interaction.* 2023, pp. 527–536.

[159] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. “Meta pseudo labels”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2021, pp. 11557–11568.

[160] Jane Allyn Piliavin and Erica Siegl. “Health benefits of volunteering in the Wisconsin longitudinal study”. In: *Journal of health and social behavior* 48.4 (2007), pp. 450–464.

[161] Ahmed Hussain Qureshi, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. “Intrinsically motivated reinforcement learning for human–robot interaction in the real-world”. In: *Neural Networks* 107 (2018), pp. 23–33.

[162] Ahmed Hussain Qureshi, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. “Show, attend and interact: Perceivable human-robot social interaction through neural attention Q-network”. In: *arXiv* (2017), pp. 1639–1645.

[163] Samantha Reig, Michal Luria, Janet Z Wang, Danielle Oltman, Elizabeth Jeanne Carter, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. “Not Some Random Agent: Multi-person interaction with a personalizing service robot”. In: *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction.* 2020, pp. 289–297.

[164] Andres Reiljan. "'Fear and loathing across party lines'(also) in Europe: Affective polarisation in European party systems". In: *European journal of political research* 59.2 (2020), pp. 376–396.

[165] Harry T Reis, W Andrew Collins, and Ellen Berscheid. "The relationship context of human behavior and development." In: *Psychological bulletin* 126.6 (2000), p. 844.

[166] Rutger Rienks and Dirk Heylen. "Dominance detection in meetings using easily obtainable features". In: *International Workshop on Machine Learning for Multimodal Interaction*. Springer. 2005, pp. 76–86.

[167] Danielle Rifinski, Hadas Erel, Adi Feiner, Guy Hoffman, and Oren Zuckerman. "Human-human-robot interaction: robotic object's responsive gestures improve interpersonal evaluation in human interaction". In: *Human–Computer Interaction* 36.4 (2021), pp. 333–359.

[168] Hannes Ritschel, Tobias Baur, and Elisabeth Andre. "Adapting a Robot's linguistic style based on socially-Aware reinforcement learning". In: *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication* 2017-Janua (2017), pp. 378–384.

[169] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Ilhan Aslan, and Elisabeth André. "Drink-o-mender: An adaptive robotic drink adviser". In: *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction*. 2018, pp. 1–8.

[170] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. "Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback". In: *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA '19. Rhodes, Greece: Association for Computing Machinery, 2019, pp. 247–255.

[171] Derek Roger and Willfried Nesshoever. "Individual differences in dyadic conversational strategies: A further study". In: *British Journal of Social Psychology* 26.3 (1987), pp. 247–255.

[172] Derek B Roger and Andrea Schumacher. "Effects of individual differences on dyadic conversational strategies." In: *Journal of Personality and Social Psychology* 45.3 (1983), p. 700.

[173] Rinat Rosenberg-Kima, Yaacov Koren, Maya Yachini, and Goren Gordon. "Human-Robot-Collaboration (HRC): social robots as teaching assistants for training activities in small groups". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 522–523.

[174] Rinat B. Rosenberg-Kima, Yaacov Koren, and Goren Gordon. "Robot-Supported Collaborative Learning (RSCL): Social Robots as Teaching Assistants for Higher Education Small Group Facilitation". In: *Frontiers in Robotics and AI* 6 (2020), p. 148.

[175] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. “Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor: 8th International Workshop on Spoken Dialog Systems”. In: 2019, pp. 247–258.

[176] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. *Active preference-based learning of reward functions.* 2017.

[177] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. “Planning for autonomous cars that leverage effects on human actions.” In: *Robotics: Science and Systems.* Vol. 2. Ann Arbor, MI, USA. 2016, pp. 1–9.

[178] Eike Schneiders, EunJeong Cheon, Jesper Kjeldskov, Matthias Rehm, and Mikael B. Skov. “Non-Dyadic Interaction: A Literature Review of 15 Years of Human-Robot Interaction Conference Publications”. In: *J. Hum.-Robot Interact.* 11.2 (2022).

[179] Sarah Sebo, Ling Liang Dong, Nicholas Chang, Michal Lewkowicz, Michael Schutzman, and Brian Scassellati. “The Influence of Robot Verbal Support on Human Team Members: Encouraging Outgroup Contributions and Suppressing Ingroup Supportive Behavior”. In: *Frontiers in Psychology* 11 (2020).

[180] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F Jung. “Robots in Groups and Teams: A Literature Review”. In: *Proc. ACM Hum.-Comput* 4.October (2020), p. 37.

[181] Ameneh Shamekhi and Timothy W. Bickmore. “A multimodal robot-driven meeting facilitation system for group decision-making sessions”. In: *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction* (2019), pp. 279–290.

[182] Garima Sharma, Shreya Ghosh, and Abhinav Dhall. “Automatic Group Level Affect and Cohesion Prediction in Videos”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019* April 2020 (2019), pp. 161–167.

[183] M.E. Shaw, R. Robbin, and J.R. Belser. *Group Dynamics: The Psychology of Small Group Behavior.* McGraw-Hill series in psychology. McGraw-Hill, 1981.

[184] Solace Shen, Petr Slovak, and Malte F. Jung. “"Stop. I See a Conflict Happening."” In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* New York, NY, USA: ACM, 2018, pp. 69–77.

[185] Hirokazu Shirado, Yoyo Tsung-Yu Hou, and Malte F Jung. “Stingy bots can improve human welfare in experimental sharing networks”. In: *Scientific Reports* 13.1 (2023), p. 17957.

[186] Elaine Schaertl Short, Katelyn Swift-Spong, Hyunju Shim, Kristi M Wisniewski, Deanah Kim Zak, Shinyi Wu, Elizabeth Zelinski, and Maja J Matarić. "Understanding social interactions with socially assistive robotics in intergenerational family groups". In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2017, pp. 236–241.

[187] Gabriel Skantze. "Predicting and Regulating Participation Equality in Human-robot Conversations: Effects of Age and Gender". In: *ACM/IEEE International Conference on Human-Robot Interaction* Part F1271 (2017), pp. 196–204.

[188] Gabriel Skantze. "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 2017, pp. 220–230.

[189] Gabriel Skantze, Samer Al Moubayed, Joakim Gustafson, Jonas Beskow, and Björn Granström. "Furhat at robotville: A robot head harvesting the thoughts of the public through multi-party dialogue". In: *International Conference on Intelligent Virtual Agents*. 2012.

[190] Gabriel Skantze, Martin Johansson, and Jonas Beskow. "Exploring turn-taking cues in multi-party human-robot discussions about objects". In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 2015, pp. 67–74.

[191] George M Slavich. "Social safety theory: a biologically based evolutionary perspective on life stress, health, and behavior". In: *Annual review of clinical psychology* 16 (2020), p. 265.

[192] Sebastian Strauß, Isis Tunnigkeit, Julia Eberle, Leonie vom Bovert, Arlind Avdullahu, Marcel Schmittchen, and Nikol Rummel. "Differential Effects of a Script and a Group Awareness Tool on the Acquisition of Collaboration Skills". In: *Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning-CSCL 2023, pp. 75-82*. International Society of the Learning Sciences. 2023.

[193] Sarah Strohkorb, Ethan Fukuto, Natalie Warren, Charles Taylor, Bobby Berry, and Brian Scassellati. "Improving human-human collaboration between children with a social robot". In: *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016* (2016), pp. 551–556.

[194] Sarah Strohkorb Sebo, Ling Liang Dong, Nicholas Chang, and Brian Scassellati. "Strategies for the Inclusion of Human Members within Human-Robot Teams". In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: ACM, 2020, pp. 309–317.

[195] Sarah Strohkorb Sebo, Margaret Traeger, Malte F. Jung, and Brian Scassellati. “The Ripple Effects of Vulnerability: The Effects of a Robot’s Vulnerable Behavior on Trust in Human-Robot Teams”. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI ’18* February (2018), pp. 178–186.

[196] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* 2018.

[197] Lei Tai, Jingwei Zhang, Ming Liu, and Wolfram Burgard. “Socially compliant navigation through raw depth inputs with generative adversarial imitation learning”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA).* IEEE. 2018, pp. 1111–1117.

[198] Henri Tajfel, Michael G Billig, Robert P Bundy, and Claude Flament. “Social categorization and intergroup behaviour”. In: *European journal of social psychology* 1.2 (1971), pp. 149–178.

[199] Hamish Tennent, Solace Shen, and Malte Jung. “Micbot: a peripheral robotic object to shape conversational dynamics and team performance”. In: *2019 14th ACM/ IEEE International Conference on Human-Robot Interaction (HRI).* IEEE. 2019, pp. 133–142.

[200] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. “Computational human-robot interaction”. In: *Foundations and Trends® in Robotics* 4.2-3 (2016), pp. 105–223.

[201] Andrea L. Thomaz and Cynthia Breazeal. “Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance”. In: *Proceedings of the National Conference on Artificial Intelligence* 1 (2006), pp. 1000–1005.

[202] Sebastian Thrun, Maren Bennewitz, Wolfram Burgard, Armin B Cremers, Frank Dellaert, Dieter Fox, Dirk Hähnel, Charles Rosenberg, Nicholas Roy, Jamieson Schulte, et al. “MINERVA: A tour-guide robot that learns”. In: *Annual Conference on Artificial Intelligence.* Springer. 1999, pp. 14–26.

[203] Leimin Tian and Sharon Oviatt. “A taxonomy of social errors in human-robot interaction”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 10.2 (2021), pp. 1–32.

[204] Margaret L Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A Christakis. “Vulnerable robots positively shape human conversational dynamics in a human–robot team”. In: *Proceedings of the National Academy of Sciences* 117.12 (2020), pp. 6370–6375.

[205] Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, Michelangelo Fiore, et al. “Spencer: A socially aware service robot for passenger guidance and help in busy airports”. In: *Field and service robotics.* Springer. 2016, pp. 607–622.

[206] Dina Utami, Timothy W. Bickmore, and Louis J. Kruger. "A robotic couples counselor for promoting positive communication". In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 2017, pp. 248–255.

[207] George E Vaillant. *Aging well: Surprising guideposts to a happier life from the landmark study of adult development.* Hachette UK, 2008.

[208] Sanne Van Waveren, Christian Pek, Jana Tumova, and Iolanda Leite. "Correct me if I'm wrong: Using non-experts to repair reinforcement learning policies". In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2022, pp. 493–501.

[209] Marynel Vázquez, Elizabeth J Carter, Jo Ana Vaz, Jodi Forlizzi, Aaron Steinfeld, and Scott E Hudson. "Social group interactions in a role-playing game". In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. 2015, pp. 9–10.

[210] Marynel Vázquez, Elizabeth J. Carter, Braden McDorman, Jodi Forlizzi, Aaron Steinfeld, and Scott E. Hudson. "Towards Robot Autonomy in Group Conversations". In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17* (2017), pp. 42–52.

[211] Marynel Vázquez, Alexander May, Aaron Steinfeld, and Wei-Hsuan Chen. "A deceptive robot referee in a multiplayer gaming environment". In: *2011 international conference on collaboration technologies and systems (CTS)*. IEEE. 2011, pp. 204–211.

[212] Marynel Vázquez, Aaron Steinfeld, and Scott E Hudson. "Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach". In: *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2016, pp. 36–43.

[213] Freydis Vogel, Christof Wecker, Ingo Kollar, and Frank Fischer. "Socio-cognitive scaffolding with computer-supported collaboration scripts: A meta-analysis". In: *Educational Psychology Review* 29 (2017), pp. 477–511.

[214] Robert Waldinger and Marc Schulz. *The Good Life: Lessons from the World's Longest Scientific Study of Happiness*. Simon and Schuster, 2023.

[215] Nigel Ward and Wataru Tsukahara. "Prosodic features which cue back-channel responses in English and Japanese". In: *Journal of Pragmatics* 32.8 (2000), pp. 1177–1207.

[216] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. "How to shape the humor of a robot-social behavior adaptation based on reinforcement learning". In: *Proceedings of the 20th ACM international conference on multimodal interaction*. 2018, pp. 154–162.

[217] Kipling D Williams. "Dyads can be groups (and often are)". In: *Small Group Research* 41.2 (2010), pp. 268–274.

[218] Kipling D Williams. "Ostracism". In: *Annu. Rev. Psychol.* 58 (2007), pp. 425–452.

[219] Pieter Wolfert, Gustav Eje Henter, and Tony Belpaeme. "Exploring the Effectiveness of Evaluation Practices for Computer-Generated Nonverbal Behaviour". In: *Applied Sciences* 14.4 (2024), p. 1460.

[220] Stephen C Wright, Arthur Aron, Tracy McLaughlin-Volpe, and Stacy A Ropp. "The extended contact effect: Knowledge of cross-group friendships and prejudice." In: *Journal of Personality and Social psychology* 73.1 (1997), p. 73.

[221] Yuto Yamaji, Taisuke Miyake, Yuta Yoshiike, P Ravindra S De Silva, and Michio Okada. "Stb: Child-dependent sociable trash box". In: *International Journal of Social Robotics* 3.4 (2011), pp. 359–370.

[222] Jing Yu, Liqi Zhu, and Alan M. Leslie. "Children's Sharing Behavior in Mini-Dictator Games: The Role of In-Group Favoritism and Theory of Mind". In: *Child Development* 87.6 (2016), pp. 1747–1757.

Part II

Included Publications