Abstract

Building Intelligent Robots for Social Regulation Therapy

Rebecca Ramnauth
2025

Throughout life, we learn the rules of social behavior by observing others, by exposure to diverse social contexts, and, in some cases, through targeted intervention. More than learning the rules and expectations on how to behave, social regulation involves the dynamic, real-time coordination of one's internal states and outward behaviors to meet those expectations. Regulation becomes challenging when internal states conflict with external demands, such as in moments of heightened emotion, sensory overload, or social ambiguity. In certain contexts (such as isolation during a global pandemic) or for some individuals (such as those with autism), social regulation can be difficult to achieve and even harder to sustain.

This dissertation positions robots as tools to support the learning of social regulation. Robots are embodied platforms and thus offer unique potential for enabling on-demand, physically co-present interactions. Although the field of robotics has traditionally focused on reliability and precision of motion to achieve physical task assistance, a growing body of literature demonstrates that humans often perceive and respond to robots as social entities. Building on this insight, we explored how robots can provide social value and assistance.

To develop such socially assistive robots, we had to overcome significant technical challenges and rethink the prevailing norms in the field. True social learning unfolds over time and requires exposure to novel real-world situations that test the relevance and adaptability of learned strategies. However, much of what we know about human-robot interaction has emerged from experimental studies in controlled laboratory or clinical environments over short timescales and typically focused on interactions between a single robot and a neurotypical adult. For robots to effectively support social regulation learning, they must operate reliably in unstructured, everyday environments; sustain long-term, repeated engagement with users of various cognitive profiles and social needs; adapt to evolving user behavior and progress; and respond in ways that are not only effective, but also socially appropriate and safe. Every component of this requires overcoming significant computational and non-computational challenges.

Across five core studies presented in this dissertation, we describe our design, de-

velopment, and deployment of robots that achieve this. While establishing feasibility is a necessary first step in ensuring that a robot operates safely, consistently, and acceptably, our work also examines whether these robots yield meaningful therapeutic outcomes. All experiments were conducted outside of laboratory settings, involved interactions spanning several days to a full month, and took place under challenging real-world conditions, including deployments in participants' homes during the COVID-19 lockdown. Each study was carefully designed to meet the needs of a highly specialized and protected user population. Collectively, these studies demonstrate the value of robots for encouraging a wide range of regulation skills, including attention sharing, turn-taking, conversational reciprocity, resiliency to interruptions, deep breathing, and emotional de-escalation.

This dissertation presents the first robots developed specifically for adults with autism. It includes one of the only robotic studies to demonstrate continuous learning progression linked to clinical measures of therapeutic efficacy. In addition, it includes the first use of foundation models to deliver unscripted and improvised therapy. It also presents the first robot to address behavioral de-escalation in public spaces while remaining agnostic to users' age or diagnostic profile.

Building Intelligent Robots for Social Regulation Therapy

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

By
Rebecca Ramnauth

Dissertation Director: Brian Scassellati

August 2025

Copyright © 2025 by Rebecca Ramnauth
All rights reserved.

For Regina—it is God's mystery in you that started this.

Acknowledgments

First and foremost, I thank my advisor, Brian Scassellati. Scaz has the unique gift of championing ideas that might be considered too ambitious, too eccentric, and too controversial anywhere else. More than that, he helps manifest those ideas in the most fun and creative ways. He has fundamentally shaped the kinds of questions I now ask and the way I go about finding the answers. I thank him for his guidance, mentorship, and belief in me from the very start.

I also thank the other members of my committee. Frederick Shic has generously shared his clinical insight and broader wisdom over the years, always encouraging me to think more expansively about the reach and impact of my work. Marynel Vázquez and Tesca Fitzgerald have modeled what it means to bring sincere joy to mentorship, a quality I hope to carry forward.

I thank the many groups and individuals with whom I have had the honor of collaborating. I am particularly indebted to Dražen Brščić and Nicole Salomons for their valuable feedback and support throughout these years. I also thank my colleagues at the Yale Social Robotics Lab, Interactive Machines Group, Brain Function Lab (Joy Hirsch), Child Study Center (Fred Volkmar and Pam Ventola), and Human Nature Lab (Nicholas Christakis), along with the many students and team members in these groups. It was a pleasure to learn alongside you all.

I thank the institutions beyond Yale that have meaningfully shaped my research journey, including the New York City Department of Education Public Schools and DIIT, the OECD (AI), the National Academies of Sciences, and the Ford Foundation community. Their engagement has been both motivating and instrumental in expanding the scope of this research. I also gratefully acknowledge fellowship support from the National Science Foundation (NSF) and the National Academies of Sciences, Engineering, and Medicine (NASEM).

I especially thank the many individuals and families who have participated in my research projects. I thank you for your time, your trust, and your generosity. You are at the heart of the studies presented here and I am grateful for the opportunity

to learn from you.

To my parents and sisters, thank you for everything. Mom and Dad, you provided me with sanctuary in every sense of the word; I have a sense of “home” and what it means to be loved. I know how rare that is and I do not take it for granted.

Contents

List of Abbreviations

All terms are presented in their expanded form as they first appear in each chapter. This reference guide is provided to facilitate easy lookup of relevant terms throughout the document. The Introduction (Chapter 1) and Conclusion (Chapter 10) are excluded, as no new abbreviations are introduced in those sections.

	Expanded Form	**Appears In**
HRI	Human Robot Interaction	Ch. 2, 3, 4
SAR	Socially Assistive Robot(ic)s	Ch. 2, 3, 4, 8, 9
SIA	Socially Interactive Agent	Ch. 2

	Clinical and Diagnostic Terms	
ABA	Applied Behavior Analysis	Ch. 2, 3, 8
ADI-R	Autism Diagnostic Interview–Revised	Ch. 3, 5
ADOS	Autism Diagnostic Observation Schedule	Ch. 2, 3, 5
ADHD	Attention Deficit Hyperactivity Disorder	Ch. 8, 9
AQ	Autism Quotient	Ch. 6, 8
ASD	Autism Spectrum Disorder	Ch. 2, 3, 5, 6, 8, 9
CARS	Childhood Autism Rating Scale	Ch. 3
CBD	Cognitive Behavioral Therapy	Ch. 3
CBE	Clinical Best Estimate	Ch. 5
CDD	Childhood Disintegrative Disorder	Ch. 3
DAS	Differential Ability Scales	Ch. 5
DSM	Diagnostic and Statistical Manual of Mental Disorders	Ch. 3, 5
GCA	General Conceptual Ability	Ch. 5
ID	Intellectual Disability	Ch. 8, 9
IQ	Intelligence Quotient	Ch. 5
OCD	Obsessive-Compulsive Disorder	Ch. 8
ODD	Oppositional Defiant Disorder	Ch. 9

PDD-NOS	Pervasive Developmental Disorder–Not Otherwise Specified	Ch. 3
SCQ	Social Communication Questionnaire	Ch. 3
SRS	Social Responsiveness Scale	Ch. 2, 3

	Systems and Technologies	
API	Application Programming Interface	Ch. 3–9
BERT	Bidirectional Encoder Representations from Transformers	Ch. 7, 9
GPT	Generative Pretrained Transformer	Ch. 7, 8
gRPC	Google Remote Procedure Call	Ch. 4
IFTTT	"If This, Then That"	Ch. 7
ISTAR	Interruption Skill Training and Assessment Robot	Ch. 6
LLM	Large Language Model	Ch. 7, 8, 9
NLTK	Natural Language Toolkit	Ch. 7, 9
NAT	Network Address Translation	Ch. 4
RESET	Robot-Enhanced Social-Emotional Therapy	Ch. 9
(C)RL	(Constrained) Reinforcement Learning	Ch. 7
ROS	Robot Operating System	Ch. 5–9
RTC	Real-Time Communication	Ch. 4
SSL	Secure Sockets Layer	Ch. 4
TMO	Transparent Matrix Overlays	Ch. 7
TURN	Traversal Using Relays around NAT	Ch. 4
UI/UX	User Interface/Experience	Ch. 4
VADER	Valence Aware Dictionary and sEntiment Reasoner	Ch. 7

	Study Design and Context	
COVID(-19)	Coronavirus Disease (2019)	Ch. 4, 6, 8
CRASAR	Center for Robot-Assisted Search	Ch. 4
DDL	Digital Dream Labs	Ch. 4
DOE	Department of Education (United States)	Ch. 9
IEP	Individualized Education Program	Ch. 9
ICT	Integrated Co-Teaching	Ch. 9
IRB	Institutional Review Board	Ch. 4–9
K–*N*	Kindergarten to Grade *N*	Ch. 9

NSF	National Science Foundation (United States)	Ch. 3
RCT	Randomized Controlled Trial	Ch. 3
WHO	World Health Organization	Ch. 2, 4

	Measures	
AUC	Area Under the ROC Curve	Ch. 5
DAU	Daily Active Users	Ch. 4
DoF	Degrees of Freedom	Ch. 3
FWS	Flow Work Scale	Ch. 6
HSD	Honestly Significant Difference	Ch. 5
ICC	Intraclass Correlation Coefficient	Ch. 7, 8, 9
LTE	Long-Term Engagement	Ch. 2
MAU	Monthly Active Users	Ch. 4
NARS	Negative Attitudes toward Robots Scale	Ch. 2
NPV	Negative Predictive Value	Ch. 5
PPV	Positive Predictive Value	Ch. 5
RoSAS	Robotic Social Attributes Scale	Ch. 2, 6
ROC	Receiver Operating Characteristic	Ch. 5
WAU	Weekly Active Users	Ch. 4

List of Figures

Chapter 1

Introduction

In moments of distress or dysfunction, regulation often begins not with explanation, but with simple behaviors: a pause before interrupting, a moment of sustained eye contact, a deep breath, a quick scan around the room. Humans employ these behaviors, sometimes without conscious awareness, to achieve emotional equilibrium, reinforce social norms and expectations, or maintain connection [9]. As micro-interventions, these behaviors preserve our social coherence, reinforce our psychological autonomy and resilience, and uphold our overall well-being [10].

Sustained regulation, however, relies on strategies that are *learned over time*—either implicitly, through repeated exposure to diverse social situations [11], or explicitly, through therapeutic instruction, structured support, or reflective practice [12,13]. When these mechanisms for learning are disrupted or lacking, regulation becomes difficult to achieve and harder still to maintain. For instance, a child who consistently observes others managing frustration effectively during peer conflicts may gradually internalize those strategies and apply them in similar situations. In contrast, a child who lacks this exposure, receives little explicit coaching, or has a neurodevelopmental condition that complicates learning, may struggle to develop comparable regulation strategies on their own.

We further observe this in diverse contexts and populations, for example, a young child struggling to navigate peer pressure and emotional volatility without mature coping tools [14]; an adult with autism who adapts to the cognitive demands of nuanced social interpretation in real time [15]; a senior with progressive dementia facing disorientation and identity loss [16]; and a caregiver operating under chronic stress with limited time for emotional recovery [17]. In each case, the capacity to stay regulated first depends on learning reliable strategies and then on being able to access and effectively deploy them when needed.

1.1 Potential of Robotics for Social Regulation

Robots hold significant potential as tools for supporting human social and cognitive growth by improving access to on-demand, personalized, socially situated, and physically co-present interventions [18]. Where the field of robotics has traditionally focused on the reliability and precision of motion to achieve functional task assistance, socially assistive robotics (SAR) has explored how robots can provide *social* value and assistance to people [19]. For example, SAR research has shown increased engagement, improved attention regulation, and more appropriate social behavior such as joint attention and spontaneous imitation when robots are part of the interaction [20, 21].

The significant advances in understanding social interactions between humans and robots have predominantly emerged from experimental studies in controlled laboratory or clinical environments, typically over short timescales and focused on dyadic interactions between a single human—most often a neurotypical adult—and a single robot (as reviewed in [22]). Although such controlled studies allow researchers to isolate specific interaction parameters, these approaches fail to capture the complexity, sustainability, and contextual relevance of long-term use in the real world. In extended interactions, users are habituated to novelty, expectations evolve, and the utility of a system is increasingly judged by its ability to provide meaningful, contextually appropriate support. Meeting these evolving expectations places new technical and interactional demands on SARs. These systems must be resilient to environmental variability, operate reliably in dynamic real-world settings, interpret and respond to human social signals, function autonomously without the supervision of the researcher, and sustain relevant support for individuals over time.

These mirror the conditions necessary for sustained regulation. True social learning does not unfold within a single 30-minute to an hour-long study session; instead, it develops over days to months, through exposure to novel, real-world situations that test the ongoing relevance and adaptability of learned strategies. It occurs beyond designated "therapy time," without constant supervision or reinforcement, and accommodates diverse cognitive profiles as well as evolving user behaviors and needs.

This dissertation presents the design, development, and deployment of SAR systems that support sustained social regulation. While much of the literature focuses on *emotional regulation*—the processes by which individuals modulate their emotional states to meet situational demands—this work adopts the term *social regulation* to emphasize the dynamic, interpersonal nature of regulation within social interactions.

In essence, it is the *social learning of emotion regulation*. We build upon recognized definitions of emotional regulation (e.g. [12,23]), but highlight regulation as a socially situated and interactionally contingent process. Learning to regulate involves, for example, developing skills for managing frustration when interrupted during a focused task; negotiating attention and turn-taking during cooperative activities; appropriately initiating or inhibiting actions toward others when emotionally overwhelmed; and adapting behavior to appropriately respond to the dynamic cues of others and the surrounding environment.

To delineate the goals of our work, we distinguish these related concepts as follows:

> **Emotion Regulation (General)**. This is the broad ability to monitor, evaluate, and modify one's emotional reactions across contexts [12,23]. It may occur in solitude (e.g., calming oneself during private stress) or in nonsocial situations (e.g., managing frustration while solving a math problem). At its core, it is an intrapersonal process.
>
> **Emotion Regulation in Social Situations**. This is the regulation of emotions specifically in the presence of others. While still focused on internal management (e.g., not crying during an argument, not showing visible anger in a meeting), the strategies are constrained by social context and expectations [12,14,24].
>
> **Social Learning of Emotion Regulation Skills (Social Regulation)**. More than regulating in social contexts, this refers to how regulation skills are acquired and refined through social interaction. It involves observation, modeling, feedback, and practice within reciprocal exchanges, where success depends not only on internal balance but also on social appropriateness, coordination, and relationship maintenance. We provide opportunities for the social learning of emotional regulation skills by building and deploying *social* robots.

These distinctions are critical to this dissertation, as they emphasize that regulation skills are learned and sustained through interaction (e.g., turn-taking, reciprocity, responsiveness). They also clarify that our SAR studies are not merely aimed at teaching private coping strategies, but at fostering socially embedded skillsets. Finally, these definitions highlight how our contributions diverge from much of the psychology literature on emotion regulation, which has predominantly examined emotional regulation as an intrapersonal process in solitary laboratory tasks (e.g., reappraisal during picture viewing; [12,25–28]). By contrast, the present work frames regulation as a socially situated, interactionally contingent skill that is learned and practiced within dynamic exchanges.

The studies compiled in this dissertation begin by examining the architectural and interactional design of robots that function with the necessary intelligence to operate autonomously, in dynamic, unstructured environments, alongside humans of diverse cognitive profiles and social needs. Then, we implement these design choices to create extended (spanning weeks or months), "in the wild" (e.g., in homes, or public schools) robot-directed interventions that support learning regulation strategies (e.g., building resilience to interruptions, mitigating social isolation during a global pandemic lockdown, or managing emotional de-escalation in a public setting) for understudied user populations (e.g., adults with autism, persons with multiple co-occurring neurodevelopmental conditions, young children receiving specialized education).

1.2 Why This Work is Challenging

Developing these robots entails a range of computational and noncomputational challenges. In the following, we list a few areas where both types of challenge converge.

1. **Heterogeneity of User Profiles.** Humans differ widely in their developmental trajectories, interaction styles, personalities, preferences, and cognitive functioning—especially within highly heterogeneous populations such as individuals with autism. This variability presents both a design and modeling challenge: robots must operate flexibly without relying on uniform behavioral baselines or one-size-fits-all interaction patterns. Our approach to this is reflected in iterative, co-design methodologies, through which we collaborate directly with specialized populations to understand their needs and inform design objectives (e.g., [29–31]). In practice, we developed systems that operate without requiring individualized pre-training, instead adapting though behavior trees or symbolic overlays that adjust to observed user behavior in real-time (e.g., [32]), robust default strategies to function reasonably across a wide range of behaviors (e.g., [32–34]), and guardrails that constrain generative outputs to ensure safety and appropriateness in novel, unanticipated scenarios (e.g., [32, 33]).

2. **Implicit Nature of Regulation Skills.** Many social regulation behaviors (e.g., eye contact, turn-taking) are learned implicitly and vary contextually. Because these behaviors are not governed by fixed rules and are rarely taught through explicit instruction, they are not easily scripted or pre-programmed.

Systems that rely on rigid rule-based approaches can produce interactions that are brittle, unnatural, or short-lived. To address this, our robots must first be capable of simulating or modeling the target behavior, either to convey its appropriate expression or to effectively prompt it in users (e.g., [29, 30, 35]). They must also recognize when user behaviors align with desired outcomes in real time (e.g., [30,31]), and crucially, infer when and how to respond, reinforce, or give feedback to support continued learning and engagement (e.g., [30,31,35]).

3. **Invisible Internal States**. Social regulation depends on internal emotional and cognitive states (e.g., frustration, anxiety, attention) that are not directly observable. Inference must occur through noisy proxies like gaze, latency, speech patterns, or physiological data—each with limited reliability and especially fragile under real-world or individual user variation. While extensive research has focused on developing reliable off-the-shelf models for automated user behavior detection, we frequently encountered limitations when applying these models to our specific user populations and deployment contexts. For example, gaze estimation models trained on neurotypical adults often failed to generalize to children with autism, whose gaze behavior may be atypical (e.g., [35]). In-home detection systems struggled with false positives due to human-like faces on televisions, toys, or images (e.g., [29,30]). Similarly, speech transcription became unreliable when the robot must distinguish between user-directed speech and ambient dialogue from other people or media sources (e.g., [34]). In the absence of reliable off-the-shelf perception models, our systems involve hybrid approaches that combine lightweight heuristics, contextual rules, and adaptive thresholds tailored to the deployment environment (e.g., [29–31]). Rather than always assuming high-confidence detection, we designed interactions to accommodate inevitable ambiguity—enabling the robot to use probabilistic reasoning or strategic deferral to deliver relevant responses even when input signals are noisy, incomplete, or misleading.

4. **Temporal Dynamics of Learning**. Social regulation unfolds gradually over weeks or months through repeated exposure, not during brief, single-session interventions. This extended timescale makes it difficult to isolate causal effects, assess short-term progress, or capture moment-to-moment learning inflections. Accounting for slow and nonlinear learning trajectories contrasts the brief, highly structured sessions typical of most robotic interventions. As we reviewed in Chapter 2, the field remains focused on proof-of-concept studies

and feasibility pilots, which tend to prioritize novelty, mere exposure effects, or initial engagement. In order to support the kind of long-term learning required for meaningful gains in social regulation, robots must sustain user engagement over time, move beyond scripted, reactive behaviors toward more proactive and generative interactions, and detect gradual patterns of change in situ. By deploying systems to operate for multiple days or weeks at a time, we create a rich testbed for exploring methods to detect user progress in situ and sustain long-term use (e.g., [29–31, 35, 36]).

5. **Social Risk and Sensitivity**. Intervening in emotional or interpersonal challenges is socially high-risk. A robot that offers feedback too early, misreads intent, appears overly prescriptive, or oversteps personal boundaries risks undermining users' trust, exacerbating stress, or causing lasting harm. Determining *when* and *how* to respond—not just *what* to say—requires fine-grained, real-time modeling of turn-taking, user readiness, and attention. To address this, robots must be able to infer latent social cues and strategically adjust or delay their interventions until the context is appropriate. Discerning *appropriateness* is the core challenge: it is rarely a discrete output, and more often an emergent property shaped by sensitivity to timing, social norms, expectations, intent, and the ongoing calibration of trust. Our efforts to formalize what constitutes socially appropriate behaviors—to ultimately enable robots to act autonomously within those bounds—have resulted in several theoretical and applied frameworks (e.g., [32–34]).

   While much attention in SAR design is devoted to onboarding and engagement, the offboarding process (how a robot exits the user's life after the intervention ends) is equally important. To holistically address the social risks that shape interaction and system design, we must recognize that relationships formed with robots, particularly those embedded in users' homes over extended periods, can carry significant emotional weight. In our work, we treat the entire deployment pipeline—including the introduction of the robot, its physical setup, in-situ troubleshooting or maintenance, exit strategies, and offboarding—as a series of essential design considerations (e.g., [29–31]).

6. **Evaluation of Therapeutic Outcomes**. Regulation is a slow and contextually embedded process, and few standardized measures exist for autonomous, unsupervised learning in social domains. All of the work presented in this dissertation features deployments outside controlled laboratory or clinical en-

vironments, occurring instead in users' everyday spaces, where interactions are minimally constrained and designed to be highly adaptable and personalized. As a result, defining a reasonable control condition—against which to evaluate both the impact of the robot-assisted intervention and baseline behavior in its absence—is often difficult or infeasible.

In addition to these experimental constraints, measuring long-term transfer, generalization, and internalization of skills remains a challenge—both conceptually and methodologically. When systems are deployed for extended, repeated interactions, they can generate hundreds of hours of interaction data (as shown in several of our studies, [29–31, 35]), making manual analysis labor-intensive, error-prone, or altogether impractical. To address these limitations, we explore methods to detect behavioral change through lightweight or passive observation, focusing on real-time processing from the system's point of view.

In summary, for robots to effectively support social regulation learning they must operate reliably in unstructured, everyday environments; sustain long-term, repeated engagement with users of various cognitive profiles and social needs; adapt to evolving user behavior and progress; and respond in ways that are not only effective, but also socially appropriate and safe. While establishing feasibility—ensuring a system operates safely, consistently, and acceptably—is a necessary first step, our work must further assess whether these systems yield meaningful therapeutic outcomes.

1.3 Dissertation Structure & Contributions

The central aim of this dissertation is:

> How can we design robotic interventions that support social regulation learning, and what interactional, technical, and contextual factors enable their effective deployment?

We begin by critically examining how the field has approached extended human-robot interactions (**Chapter 2**). In this review of 120 studies, we operationalize how the field currently defines "long-term" engagement and how user outcomes are measured. This chapter highlights opportunities to expand the design scope of SAR systems, improve their readiness for real-world deployment, and improve methodological consistency across studies. These findings inform and motivate the evaluation strategies adopted in this dissertation.

We then conduct a large-scale review of over 300 studies involving the use of robots in interventions for Autism Spectrum Disorder (ASD)—not only because ASD has been a popular focus of SAR research, but also because it presents a uniquely rich testbed for studying the mechanisms of productive social learning (**Chapter 3**). Core diagnostic features of autism—such as challenges in social communication, emotional regulation, and adaptive behavior—align closely with the areas where SARs are hypothesized to provide therapeutic benefit. As such, the ASD literature offers critical insights into the potential and limitations of SAR-based interventions. In our review, we identify foundational trends, common design assumptions, proposed robot-led pedagogies for teaching valued social skills, and key research gaps that inform the broader aims of this dissertation.

Chapters 4–9 present a series of human-subject experiments, each contributing to the design, development, and deployment of a robot-based intervention. These studies aim to evaluate both the feasibility of system operation and its therapeutic impact. All experiments were conducted under challenging conditions, including during the COVID-19 pandemic lockdown, and were designed to meet the needs of a highly specialized and protected user population. Moreover, these studies demonstrate the value of SARs for encouraging a wide range of regulation skills, including attention sharing, turn-taking, conversational reciprocity, interruption tolerance, deep breathing, and de-escalation. They also underscore the importance of architectural flexibility, real-time adaptability, and socially aware design constraints for enabling long-term, autonomous operation with vulnerable users in real-world environments. The experiments presented here include the first SARs developed for adults with ASD for in-home therapy, one of the few SAR studies to demonstrate continuous learning progression tied to clinical measures of therapeutic efficacy, and the first SAR to address behavioral de-escalation in a public space while remaining agnostic to users' age and diagnostic profile.

The first experiment (**Chapter 4**) describes the development of a robot to mitigate social isolation among children during the COVID-19 pandemic. While social distancing and quarantine mandates were essential for public health, they intensified feelings of loneliness—an issue already recognized as a growing societal concern. Because children at this developmental stage acquire critical, life-long social skills through physical play, we created a system that allowed one child to remotely control and communicate through a robot located in a peer's home, allowing them to engage in physical play while being geographically separated. With over 2,000 unique users in three months, this study offered valuable insights into how robots can be deployed in unstructured, home-based environments to effectively support social connection.

While Chapter 4 examines how robots can support broad social and emotional needs, **Chapter 5** shifts the focus to how robots can support specific developmental outcomes. This chapter examines the impact of a month-long, in-home, robot-assisted intervention aimed at improving gaze behavior in children with ASD. Appropriate gaze behavior is a foundational component of early social development, a prerequisite for more complex social skills, and a core diagnostic feature of ASD. The intervention, conducted by Scassellati et al. in 2018 [3], was a landmark study that demonstrated both the feasibility and the promise of robot-assisted interventions for ASD. Not only did it validate that such in-home systems could be deployed successfully, but it also provided evidence of meaningful developmental gains—most notably, improvements in joint attention. However, at the time, the gold standard for evaluating these outcomes was clinician-administered assessments conducted at home once at the start and once more at the end of the intervention. Although this approach yielded promising results, it left several critical questions unanswered: When did these behavioral changes emerge during the intervention? Were they gradual or abrupt? Consistent between participants or highly individualized?

Understanding the timing of behavioral change has important implications for the future of autonomous therapeutic systems. If we can identify *when* behavioral improvements occur, it may be possible to develop systems capable of autonomously detecting those inflection points—recognizing, in real time, when they effectively support users. To achieve that goal, we needed to revisit the computational methods for automatically extracting and interpreting behavioral change. In this chapter, we address these open questions: Was the SAR-based therapy effective? Did it lead to measurable behavioral improvements? Can behavioral change be automatically and accurately detected from interaction data? When exactly did these changes emerge? More broadly, what do these patterns reveal about ASD and the design of robot-based interventions for such a uniquely heterogeneous population?

Still, despite decades of progress in ASD research, the vast majority of studies and clinical programs have focused almost exclusively on children. Although social, emotional, and functional challenges are well documented to persist and in some cases intensify, in adulthood, relatively few studies have addressed how to support *adults* with ASD. **Chapter 6** explores how SARs can support employment and workplace readiness for adults with ASD. We developed a robot-led intervention that simulated common workplace encounters, promoting role play and naturalistic social practice while integrating into participants' daily home routines. During the course of a week, the users engaged in managing unexpected social demands and developed strategies for cognitive and attentional regulation. Behavioral data and participant feedback revealed increased resilience to interruptions, positive perceptions of the robot's usefulness for supporting employment goals, and preliminary evidence of skill generalization. This study represents the first in-home SAR intervention specifically designed for adults with ASD.

As individuals with ASD transition into adulthood, the social demands they face become more complex, ambiguous, less easily scripted, and less forgiving of atypical behavior. This escalation in social complexity imposes greater demands on the social intelligence and adaptive capabilities of SARs intended to model or scaffold appropriate behaviors. Several of our intermediate studies (e.g., [33, 34]) explore how robots can discern when to initiate interaction by assessing social appropriateness—not merely detecting if a human user is present, but assessing whether it is contextually suitable and productive to engage. These considerations are critical for the success of later interventions, where timing, context, and user readiness shape engagement quality and outcomes.[1]

As our work with robot-assisted interventions for adults with ASD progresses, we turn to conversational skills—particularly small talk—as a critical yet undersupported domain tied to real-world outcomes. From dating to job interviews, making new friends or simply chatting with the cashier at checkout, small talk plays a key role in social integration and opportunity access, yet remains especially challenging for individuals with ASD. To address this, we explore the integration of large language models (LLMs) into SARs to support natural unscripted conversation. However, deploying an LLM-driven system for long-term, autonomous, unsupervised operation with vulnerable users in their homes requires robust safety mechanisms to ensure appropriate behavior. **Chapter 7** introduces a framework for implementing behavioral guarantees in SARs relying on foundation models, establishing safeguards that inform the development of autonomous interventions in later chapters. Chapter 7 also demonstrates the practical application of this framework to enable robots to engage in naturalistic small talk. We position small talk as a compelling frontier that reveals both the promise and complexity of deploying SARs driven by foundation models in socially sensitive contexts. Finally, **Chapter 8** presents the design, development, and deployment of a SAR-based small talk training program for adults with ASD. Com-

[1] While several formative studies contributed to the development of the systems and insights presented in this dissertation, not all are included as dedicated chapters. These supporting studies are cited where they were applied or relevant to preserve a clear narrative arc and focus on the most consequential deployments.

pared to our previous deployments, this study introduces novel interventional design considerations, including a shift beyond rote skill rehearsal toward delivering useful, personalized feedback—an especially nuanced challenge, as social-skills feedback is often deeply personal and closely tied to identity.

In **Chapter 9**, we demonstrate our guardrail mechanism in a new application area: enabling safe and effective SAR intervention to support students experiencing heightened emotional states, sensory overload, or difficulties with self-regulation in traditional classroom settings. While many schools have introduced de-escalation or sensory rooms to support these needs, their effectiveness is often limited by the wide range of student profiles and constrained staff availability. To address this, we developed a SAR to improve students' self-regulation skills within a school's existing de-escalation space. This chapter details the co-design process, iterative development, and final system architecture. Following a fully autonomous, month-long deployment in an elementary school, we evaluated the robot's usability and impact. Findings indicate that the system integrated seamlessly into the school routine, improved de-escalation efficiency, facilitated smoother transitions back to learning environments, and produced sustained positive effects months beyond the deployment period.

In **Chapter 10**, we conclude the dissertation with a summary of the work presented. We discuss key contributions and broader implications, along with directions for future research. In all, the core contributions of this dissertation are as follows.

1. **A cross-domain analysis of SAR studies** demonstrating extended interactions with SARs, operationalizing definitions of "long-term" deployment and user outcome measures.

2. **A comprehensive review of more than 300 studies** on robot-assisted interventions for individuals with ASD, identifying foundational trends, design assumptions, and research gaps in the field. We conclude this review by proposing **a consolidated hypothesis and a theoretical foundation** for why robots may be uniquely effective tools for therapy, based on empirical findings and psychological frameworks.

3. **Design, development, and deployment of multiple SAR systems** targeting social regulation skills (e.g., attention sharing, turn-taking, conversational reciprocity, interruption tolerance, deep breathing, and emotional de-escalation) across diverse populations (e.g., adults with ASD, individuals with multiple co-occurring neurodevelopmental conditions, elementary-aged children in spe-

cialized education programs) and settings (e.g., in-home deployments during pandemic lockdown periods and public school setting).

4. **The first SAR systems designed specifically for adults with ASD**, addressing an underrepresented population and targeting skill areas that are largely overlooked in both SAR development and clinical intervention research.

5. **Introduction of a safety and behavioral guardrail framework for SARs using foundation models**, enabling ethical, unsupervised deployment in socially sensitive contexts.

6. **Empirical evidence of SAR impact beyond mere novelty and presence effects**, including sustained engagement, skill generalization, and successful integration into users' everyday routines and environments.

Chapter 2

Toward Sustained Social Interaction: A Review of Trends, Gaps, and Challenges in Long-Term HRI

Over the past two decades, the field of robotics has experienced substantial growth, marked by a notable increase in long-term human-robot interaction (HRI) studies. To enable a broad inclusion of the relevant literature, we define "long-term" as studies in which a robot interacts with the same user over at least three sessions spanning a minimum of three consecutive days. As a result of adopting this inclusive definition, this chapter synthesizes 120 long-term HRI studies conducted over the past two decades. These studies span seven key domains, including education, entertainment, and physical and mental health, offering a comprehensive view of the field's evolution and scope. From this corpus, we extract key patterns and divergences in study design, participant demographics, interaction dynamics, and evaluation methods, providing a structured overview of the current landscape of long-term HRI.[1]

This review reveals emerging trends, underlying design assumptions, proposed robot interaction strategies, and critical research gaps. Together, these insights inform the goals of this dissertation in its three core dimensions: the *design* of robots for social interaction, their technical *development*, and the contextual factors that enable their successful *deployment*. The growing emphasis on long-term real-world deployments observed in this review underscores the importance of designing robots capable of sustaining meaningful engagement over extended periods of interaction. Observed pedagogical patterns, particularly in educational and therapeutic settings, offer concrete models of how robots can scaffold learning, support social regulation, and adapt to user needs over time. At the same time, persistent gaps—such as the limited inclusion of adolescents and the relative scarcity of studies in school and

[1]This chapter is adapted from our published work: Matheus, K., **Ramnauth, R.**, Scassellati, B., & Salomons, N. (2025). Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations. *ACM Transactions on Human-Robot Interaction, 14*(3), 1-42. [22]. We include additional context, commentary, and analysis to support its integration into this dissertation.

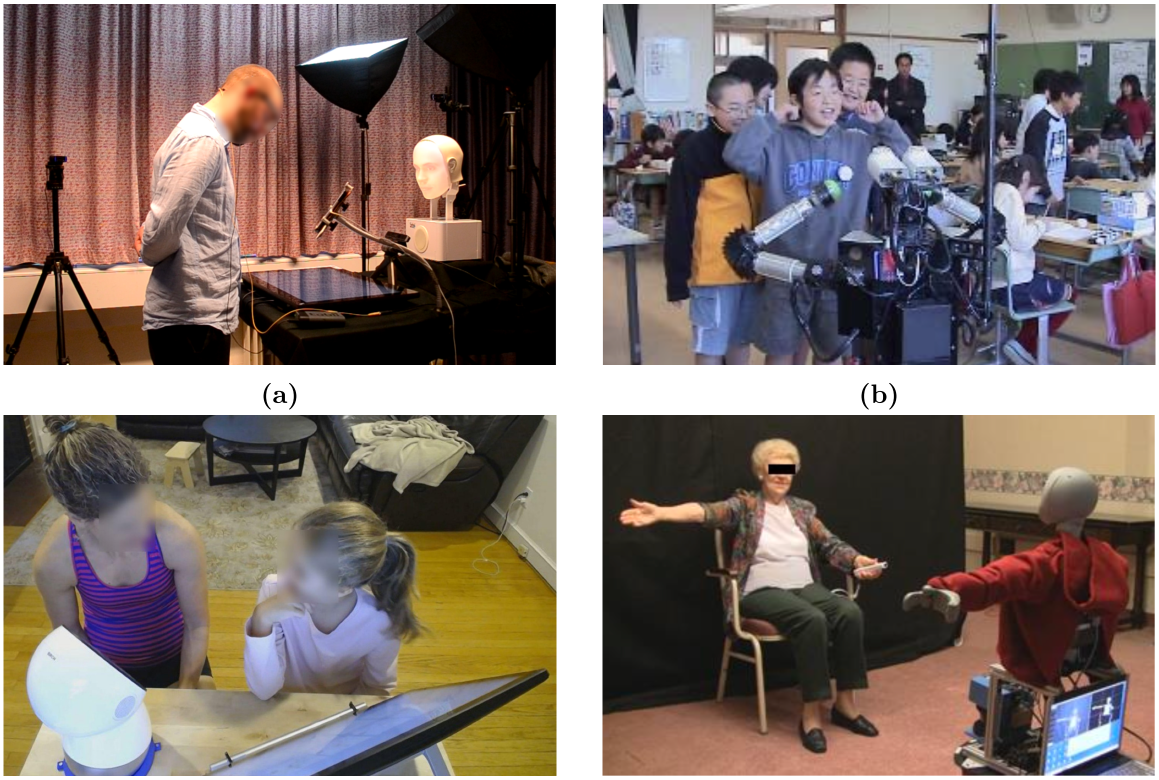
workplace contexts—highlight the need for more inclusive and context-aware interventions. This dissertation responds to these gaps by advancing robots tailored for underrepresented populations and settings, while proposing new methods to support sustained, socially meaningful interaction. In doing so, it builds on and extends the trajectory of long-term HRI research.

2.1 Introduction

During the past two decades, the field of social robotics has undergone remarkable growth, accompanied by a surge in studies that examine long-term human-robot interactions that unfold over multiple days, weeks, or even months. A *social robot* possesses features and capabilities that enable it to interact with humans in ways that resemble social interactions between people [19]. Such robots can exhibit behaviors such as recognizing and expressing emotions [37], understanding and generating natural language [38], adapting to different social contexts [39], and even demonstrating a degree of empathy [40]. For many, the goal of social robotic study is to support a future where robots are not only present in individual laboratory sessions, but are integrated *longitudinally* into daily lives.

By our accounts, the number of HRI research papers examining the longitudinal use of social robots has tripled from before 2013 to 2023 (Section 2.4). Bajones et al. [41] have highlighted this transformation by noting that whereas the "burning question in HRI studies" was once "how many participants do I need?", it is now "how long should my study run for?" Such a shift reflects the field's progression beyond initial experimentation in social robotics and toward a deeper understanding of how robots can be effectively deployed in real-world applications. In a future where robots are integrated into our homes, schools, offices, and medical facilities, it becomes increasingly essential to research the dynamics of long-term interactions spanning days, weeks, months, and even years.

From a research perspective, the study of repeated human-robot interactions offers several distinct advantages over single-session studies. First, many forms of human-robot interaction require longitudinal engagement to achieve meaningful impact—particularly in applications such as tutoring, training, or therapy, where robots are intended to support skill development or behavioral change. Second, as in human-human relationships, the dynamics of human-robot relationships evolve over time. The key benefits of long-term study are understanding how to foster healthy and resilient HRI relationships, how to personalize interactions with individual users, and



(a) (b) (c) (d)

Figure 2.1: Illustrative Cases of Long-Term Robot Deployment Across Domains. Four long-term robotic interaction studies from our corpus, illustrating the diversity of application domains and interaction characteristics analyzed in this review: (a) a study on the persistence of first impressions with a Furhat robot [1], (b) a Robovie robot engaging children in a classroom setting [2], (c) a robot delivering social skills training to children with Autism Spectrum Disorder [3], and (d) a robot designed to motivate physical exercise among older adults [4].

which relational models are most effective in different contexts. Finally, long-term deployments are essential for investigating practical factors such as usage patterns, drop-off rates, and strategies for sustaining engagement [42–44]. These insights are crucial to understanding the effective integration of robots into daily life and society. Only by studying HRI over longer periods can researchers observe how users adopt robotic systems and what factors contribute to successful and lasting adoption. Section 2.2.2 provides a deeper discussion of these benefits.

Although the benefits of long-term HRI research are substantial, the pursuit of such studies often demands significantly more time, resources, and effort from researchers [45]. Unlike single-session experiments, long-term studies typically involve repeated interactions across multiple sessions. Although they can be conducted in controlled laboratory environments, they are less commonly situated there due to practical constraints such as limited space, scheduling difficulties, and challenges in maintaining participant availability. As a result, long-term studies are more likely to be carried out in real world settings such as homes, schools, or clinics, where robots are expected to operate "in the wild." These scenarios demand robots that are robust and autonomous, capable of reliably functioning in diverse contexts and at varying times of the day. Researchers must account not only for how the robot behaves during scripted interactions but also for how it operates during unsupervised moments throughout the full deployment timeline. Investigations of topics such as personalization and adaptive behavior to maintain user engagement require additional layers of design and technical sophistication. In addition, long-term deployments raise concerns about data privacy, ethical oversight, and system reliability, while participant recruitment and retention over extended periods remain persistent logistical challenges. A more detailed discussion of the challenges and trade-offs involved in long-term HRI research is presented in Section 2.2.3.

Because the nature of long-term HRI studies presents substantial challenges, it is vital that researchers look to previous efforts for guidance, inspiration, and cautionary lessons. Therefore, we provide a comprehensive review of long-term HRI studies conducted over the past 20 years as a valuable resource for researchers. Our work fills a notable gap in the literature, as the last in-depth review on this topic [46] was conducted more than a decade ago in 2012. Building upon this prior work, our study presents a comprehensive analysis of 120 studies spanning seven major domains, such as education, entertainment, physical health, and mental health. Our analysis covers the period from 2003 to the time of analysis (April 2023), providing a comprehensive overview of the progress made in the field. To ensure inclusiveness and capture relevant trends, we have defined *long-term* in our corpus as studies deploying a robot with the same users for three or more sessions over three consecutive days (e.g., one session per day for three days, once a week for three weeks, etc.). By adopting this criterion, we aim to identify patterns and insights that can contribute to a deeper understanding of long-term HRI. Nonetheless, we also discuss alternative approaches to the definition of "long-term" in Section 2.2.1.

In our analysis, we explore patterns over time and across domains for a number of study design aspects, including longitudinal characteristics of the studies, the types and number of participants involved, study locations, defining elements of human-robot interaction, and the types of results and engagement measures employed. Figure 2.1 shows four different long-term robotic interaction studies from our corpus, which vary widely between the characteristics we analyze. For instance, studies vary from dyadic (a and d) interactions to triadic (c) and group (b) interactions. The ages of the participants in our studies ranged from children (b, c) to the elderly (d). Some studies were carried out in the wild, such as in the home (c) or schools (b); others (a and d) were conducted in laboratory environments. In addition, they span a variety of domains, including education (a and b), social skills training (c), and physical exercise (d).

By highlighting key insights and design patterns from existing studies, our aim is to equip the HRI community with the knowledge needed to collectively address persistent challenges, refine methodological approaches, and drive innovation in the field of long-term human-robot interaction. We conclude this review by outlining major gaps in the literature, identifying opportunities for future research, and offering practical guidelines to support the design and execution of long-term HRI studies.

2.2 Background

In the sections that follow, we establish a working definition of long-term within HRI, critically assess the benefits and challenges of sustained interaction research, and summarize previous review efforts in the field.

2.2.1 What is Long-Term HRI?

As this review covers the topic of long-term HRI, the first question one may ask is: what does "long-term" mean in the context of robotics? Could *any* repeated interaction with a robot be considered long-term, or is there an unspoken rule that a study must meet in order to be considered part of this category? Multiple perspectives are available in the HRI literature with no single agreed-upon approach. Many studies are organized around distinct sessions between users and robots, so one logical approach is to consider a minimum number of sessions. Kory et al. take this approach for inclusion in their review of long-term, socially interactive agents [47], using five sessions as a cutoff while also acknowledging the arbitrary nature of this choice. By this measure, Ramachandran et al.'s [48] five sessions with an adaptive robotic tutor would be considered long-term, but Donnermann et al.'s [49] three sessions or Jones and Castellano's [50] four sessions, also with adaptive robotic tutors, would not. Arguing that there is a meaningful distinction between, for example, four versus five sessions as a threshold for "long-term" may be unproductive. Moreover, if all five sessions occur within a single day rather than being distributed over time, the validity of the “long-term” designation becomes more questionable. Similarly, how should we compare a series of five 3-minute sessions across a week to five 30-minute sessions? Although the number of sessions may be an easy yardstick, it does not fully account for the depth, duration, or temporal spacing of interactions and therefore provides an incomplete picture of the full interaction story.

An alternative approach to defining “long-term” might involve calculating the total duration of the user-robot interaction over the course of a study. Few studies explicitly report this metric; for example, Afyouni et al. [51] observed an average of 2 hours and 22 minutes of free-use interaction over one week, while Kidd and Breazeal [52] reported an average usage spanning 50.6 days. However, most studies do not measure or report total interaction time, potentially due to methodological complexity or a lack of historical precedent. Even when reported, this metric does not account for the distribution of interactions over time—for example, the frequency and length of gaps between sessions. More fundamentally, the interpretation of what constitutes “long-term” may vary depending on the robot’s domain and intended function. A given duration of interaction may have very different implications for a tutoring robot compared to a game-playing or therapeutic robot, due to differences in cognitive load, task structure, and the degree of repetition involved. Thus, while the total interaction time offers a useful starting point, it also remains an incomplete measure of longitudinal engagement in HRI.

Another approach to benchmarking the definition of “long-term” in HRI is to evaluate whether the interaction has surpassed the *novelty effect* [46]. The novelty effect refers to the initial period of heightened interest or excitement when users first encounter a new technology. It is an effect that is expected to fade as familiarity sets in. In the context of social robotics, this transition is marked by a shift from surface-level curiosity to more stable, predicable patterns of user behavior. Designing studies that capture user behavior beyond this initial novelty phase allows researchers to more confidently attribute observed outcomes to sustained human-robot interaction, rather than to short-term responses driven by the appeal of a new experience. As such, passing the novelty threshold may serve as a more functionally meaningful indicator of “long-term effects” than session count or total interaction time alone.

To this point, Bajones et al. [41] proposed that the novelty phase is unlikely to subside before three weeks of repeated interaction, and thus established a study duration of 21 days to ensure their findings reflected post-novelty, long-term engagement. Previous work has attempted to establish when the novelty effect wears off [42, 46]. However, there is no fixed or universally agreed-upon timeframe given the numerous variables that can influence novelty effects. For instance, the complexity of the interaction affordances on the robot may impact how inherently engaging it is and the diversity of features to explore. Alternatively, the setting of the robot (e.g., public versus private) may impact the level of desire a user has to interact with the robot. Such variability ultimately places the responsibility on researchers to substantiate attenuation of the novelty effect in their specific use-case scenario by employing appropriate metrics and analysis, which relatively few authors in our corpus have done. Among those who have, Weiss et al. [42] noted the presence of novelty effects by observing a decline in attachment levels with a Vector robot after just two weeks out of a 30-week study. In contrast, Bodala et al. [53] reported no discernible evidence of novelty effects even after five weekly sessions with a mindfulness training robot. This finding presents a methodological puzzle: Did the study fail to elicit novelty effects from the outset, or was a five-week duration still insufficient to move beyond them?

In the absence of a universally accepted definition, this review adopts a deliberately broad approach to defining *long-term* within the context of HRI. For the purposes of our analysis, we consider studies to be *long-term* if they involve interactions with the same user in more than three sessions on at least three separate days. This threshold allows for consistency across our corpus while acknowledging the field's variability. In this framing, all long-term studies are necessarily multisession, though not all multisession studies qualify as long-term—such as those limited to only two sessions. Although we do not aim to establish a rigid definition of "long-term" HRI, one of our goals is to clarify how the field has operationalized this concept to date. In Section 2.4.1, we examine the diverse temporal characteristics represented in our corpus, including session count, number of days, and total interaction time, and offer guidance on how future research might more thoughtfully incorporate these elements.

2.2.2 Benefits of Long-Term HRI Research

Long-term HRI studies inherently require substantial time, resources, and effort. In this section, we outline three key reasons why researchers may find it valuable to invest in conducting long-term studies in human-robot interaction.

Interaction Outcomes

Several types of robot interaction and opportunities are only possible with long-term deployments. Certain robotic systems designed to teach new skills, such as tutoring, physical training, or cognitive training, require multiple sessions over an extended period to effectively support skill development (e.g., [4, 49, 54–60]). Assessing the effectiveness of these programs also requires longitudinal measurements to demonstrate the user's skill acquisition over time. Therapeutic robots, such as those that support people with mental health challenges [61] or Autism Spectrum Disorder (ASD) [62], benefit from a long-term study in similar ways. Successful therapeutic engagements are often repetitive in nature [63], and long-term interaction allows researchers to observe the effects of sustained engagement on the well-being of individuals receiving therapy.

User Perceptions and Relationships

Another key advantage of long-term HRI studies is the opportunity to examine how users' perceptions of robots, and their relationships with them, evolve over time. Kory-Westlund et al. introduce the concept of moving "beyond interaction to relationship" in long-term HRI [47], noting that just as relationships with humans can enhance learning, health, and social outcomes, sustained relationships with robots may offer similar benefits. For example, Paetzel et al. [1] found that while robot perceived competence remained consistent throughout repeated sessions, perceptions of threat and discomfort varied, suggesting a more nuanced relational dynamic. Jeong et al. [64] observed that a home-based social robot was able to build rapport and a working alliance while supporting the mental well-being of college students. The ability to foster positive, long-term relationships is particularly important for companionship-oriented robots, such as those designed for general home use or to mitigate loneliness in elder care settings [6, 8, 65–68].

Robot Adoption and Long-Term Engagement

Long-term studies also offer valuable insight into the adoption of robots in various settings. Factors such as usage patterns, drop-off rates, and strategies to sustain participation [42–44] are crucial to understanding the effective integration of robots into daily life and society. One specific area of interest is the examination of the novelty effect and its impact on user interactions with robots [69, 70]. Consequently, a key challenge for long-term HRI is designing systems that continue to engage users even after the novelty effect has subsided [6, 71]. Only by studying long-term robot deployments can researchers explore strategies and design principles that promote both adoption and sustained user engagement, ensuring that robots maintain their effectiveness.

2.2.3 Challenges of Long-Term HRI Research

Despite the many benefits of long-term HRI, conducting such studies presents several nontrivial challenges. In this section, we provide a brief overview of the key obstacles encountered by the studies in our corpus. These challenges underscore the importance of understanding the current state of the field and highlight areas where methodological refinement and community-wide support are especially needed.

Autonomous, In-the-Wild Deployments

Longer-term deployments often increase the demand for robot autonomy, as it becomes impractical to rely on a human operator throughout the study period. Although not all long-term studies face this challenge, many require the development of fully autonomous systems equipped with more complex capabilities such as robust behavioral planning, expanded sensorimotor functions, wireless connectivity, remote data logging, and streamlined remote troubleshooting.

In addition, many long-term studies take place in "in the wild" environments such as homes, schools, or hospitals. These real-world settings offer valuable opportunities to study robots in authentic contexts and often facilitate access to relevant participant populations at scale. However, conducting research in these environments also presents considerable challenges, as researchers have little control over physical or social conditions. Each home, classroom, or clinic presents a unique set of environmental variables that require robots to adapt to diverse and unpredictable conditions. For example, perception systems must handle varying levels of background noise and lighting; safety protocols must account for a wider range of scenarios; and behavior planning must remain robust across fluctuating user behaviors and contextual cues.

Recruitment and Adherence

Recruiting participants for long-term HRI studies is often significantly more complex than for single-session research. Unlike short-term studies, participants must commit to multiple sessions over an extended period, increasing the likelihood that changes in schedules, routines, or personal circumstances will interfere with participation. Additionally, depending on the study context, participants may need to welcome the robot into more private or personal spaces, such as their homes, which can further narrow the pool of willing volunteers.

The maintenance of participation also becomes more challenging with time. In short-term studies, users are generally able to remain engaged and adhere to protocols, particularly under the supervision of researchers. In contrast, long-term deployments—especially those conducted in the wild—require participants to remain engaged without continuous oversight. After the initial novelty wears off, users must be intrinsically motivated to continue interacting with the system. As such, long-term systems must offer sustained value, whether through engaging behaviors, perceived usefulness, or meaningful integration into daily life. These increased demands place a greater burden on both the system design and the participant experience.

Dynamic Content and Personalization

Developing sufficient content and robot behaviors for long-term deployments can be a substantial challenge—particularly when interactions span weeks, months, or even years. Unlike short-term studies, where limited and repetitive content may suffice, long-term HRI requires a more extensive and varied interaction repertoire. Users are unlikely to remain engaged with a robot that delivers the same or overly similar utterances and actions across sessions, making it considerably more challenging to sustain interest over time.

This challenge is compounded when considering personalization, which has been shown to improve user engagement and interaction outcomes in numerous studies [72, 73]. Personalizing content for individual users introduces additional complexity, as it demands not only a larger pool of content but also consideration of diverse user attributes such as age, preferences, skill levels, and learning styles.[2] Personalization also requires robust user modeling systems and intelligent action selection mechanisms capable of adapting to the user's evolving state over time.

In general, the design and development of interaction content for long-term studies is considerably more labor intensive and technically demanding than for single-session studies, which poses a significant barrier to scalability and widespread deployment.

Cost

Another critical consideration when designing a long-term study is the cost of the hardware systems involved. Robots are often expensive to acquire, maintain, and operate—particularly when multiple units are required to support parallel deployments with several users. In studies where only a single robot is available, the need to

[2] It is worth noting that recent advances in large language models have begun to ease some of these challenges, particularly with regard to dynamic content generation. However, the vast majority of studies in our corpus were conducted before such technologies became widely available.

sequentially run participants can dramatically extend the study duration, potentially taking months or even years to complete. Beyond the robots themselves, long-term studies often require duplicate sets of supporting equipment, including computers, cameras, networking hardware, and environmental sensors. These additional costs, both financial and logistical, can significantly constrain the scale, length, and feasibility of long-term research, particularly for studies that aim to include diverse populations or real-world contexts. As a result, the high resource demands of long-term HRI can limit not only who can conduct such studies, but also what kinds of questions can realistically be explored.

2.2.4 Prior Long-Term HRI Reviews

There have been two prior reviews specifically addressing the topic of long-term HRI [46, 47]. The first, published by Leite et al. in 2012, focused on social robots used in extended interactions and included a corpus of 24 studies [46]. The authors considered papers in which robots engaged users socially for prolonged periods, although they did not explicitly define what constituted an "extended" timeframe. The primary objective of the review was to synthesize key findings from these studies and to identify open questions for future research on long-term HRI. Notably, the review paid particular attention to how user interaction evolved beyond the novelty effect, emphasizing studies that demonstrated continued engagement over time.

In their review, Leite et al. [46] categorize the 24 studies into four application domains: healthcare and therapy, education, work environments and public spaces, and home robotics. In the domain of healthcare and therapy, the authors report generally positive outcomes, although they note that most studies were limited by small sample sizes. Their discussion of educational robots centers primarily on child-robot interactions, highlighting the critical role of the robot's form factor and behavior in shaping user engagement. For studies situated in work environments and public spaces, the authors caution against drawing broad conclusions due to the diversity of contexts and the limited number of studies available. Lastly, in the context of home robots, the review emphasizes the importance of overcoming the novelty effect to ensure sustained use. The novelty effect is described as having worn off once users become familiar with the robot and begin seeking new behaviors or experiences from the system.

There are three key differences between Leite et al.'s review and the present study. First, our review includes a broader and more recent body of work, incorporating studies published between 2013 and mid-2023. Over the past decade, the field of long-term HRI has expanded considerably, and many influential contributions are captured in our updated corpus. Second, our inclusion criteria differ substantially. Although Leite et al. include studies involving robots that interact with different users each day, such as receptionists or mall guides, we focus exclusively on studies in which the same user interacts with the robot across multiple sessions over time. This distinction allows us to differentiate between long-term *interaction* deployments, where a user experiences the robot longitudinally, and long-term *research* deployments, where the robot is deployed for an extended period but individual users may only interact with it once. Our focus is on the former, and this is what we refer to as "long-term" throughout this review. Finally, we disaggregate study characteristics related to deployment setting and application domain—for example, distinguishing between a healthcare robot used in a private home versus one deployed in a clinical facility. In contrast, the smaller number of studies available at the time of Leite et al.'s review meant that deployment context and domain were more closely correlated and not analyzed separately.

The second prior review on long-term HRI is a book chapter on long-term socially interactive agents (SIAs), published by Kory-Westlund et al. in 2022 [47]. Although their scope is broader—in that it encompasses both physical robots and digital agents—the authors dedicate a section specifically to long-term embodied robots. Their inclusion criteria define long-term interaction as involving *at least five sessions*, irrespective of session length or the time interval between the first and fifth sessions. Using this definition, they identified 67 relevant studies. Similarly to the 2012 review by Leite et al., their discussion is organized around application areas, specifically: "social robots and children," "social robots and health and wellness," and "living with consumer robots." The chapter summarizes key insights from studies in each of these domains and emphasizes the importance of relationship building over time in long-term SIA interactions.

Our review is complementary to this work, but differs in several important ways. First, we adopt more specific inclusion criteria, focusing on studies that involve *at least three sessions with the same user over a minimum of three separate days*. Second, we provide a broader analysis of long-term HRI studies, examining trends over the past two decades and categorizing them into seven key domains. Finally, we extend the analysis to include a range of study design characteristics, such as deployment setting, robot autonomy, personalization, and engagement metrics, as detailed in Section 2.3.

2.3 Review Method

In this section, we describe the method for compiling our corpus. In order to systematically review social robots deployed in long-term situations over the past two decades, we developed the following criteria for paper inclusion.

1. The study must involve social interaction between at least one robot and at least one human. We did not include studies in which the robot only had functional use with little to no social elements (e.g., [74, 75]).

2. The participants and the robot must have interacted for at least three sessions (either as required by the study or from the user's free use) over a minimum of three separate days.

3. The robot interacted with the same user for at least three sessions, and data was tracked for the user throughout the sessions. This excludes long-term studies in public spaces where the robot was interacting with a different user every time (e.g. [76, 77]), or where data was not clearly tracked for the same user (e.g. [78, 79]).

4. The robot must be or appear to be an autonomous agent to the user. Wizard-of-Oz studies are included in the analysis if the robot was presented to the participants as autonomous. We did not include telepresence studies (e.g. [80]) or studies in which the participants mainly controlled or programmed the robot (e.g. [81]). Furthermore, we did not include studies that focused solely on design, in which the user did not interact with the robot in a deployed manner (e.g. [82]).

5. The robot must be physically present (embodied) during the interaction. We included studies with multiple experimental conditions if at least one had an embodied robot. We did not include studies where the robot was only represented on a screen or as an avatar (e.g. [83]).

We started by including all papers from Leite et al. [46] that met our updated inclusion criteria. For studies published after 2012, we conducted a systematic keyword search using Google Scholar as well as proceedings from relevant journals and conferences. Our search terms included: *long-term*, *long term*, *in-home*, *in home*, *home deployment*, *repeat use*, *repeated use*, *multi-session*, *multisession*, *in the wild*, *in-the-wild*, *weeks*, *months*, and *longitudinal*, each used in combination with *robot* or *robotics*. The venues reviewed included major outlets in the HRI community: *Human-Robot Interaction (HRI)*, the *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, the *International Journal of Social Robotics (IJSR)*, the *Journal of Human-Robot Interaction (JHRI)*, *ACM Transactions on Human-Robot Interaction (THRI)*, and *Science Robotics*. The papers identified through these channels were initially screened based on their abstracts and then manually reviewed by three members of our research team to assess their eligibility according to our inclusion criteria. The earliest paper in our corpus was published in 2003 [84], and our search cutoff date was April 2023.

In total, we found 118 papers that met our criteria, representing 120 studies. The number of studies and papers differ as some papers (e.g. [85]) presented several separate studies that fit our criteria, and others had multiple papers derived from the same data which were not included in our final corpus. Our research team reviewed the papers and annotated key information from each, as outlined below. When uncertain or disagreed, the three annotators discussed a paper together to reach a consensus on which category or number best represented the situation. The following information was extracted from each paper:

Temporal Qualities:

- **Year**: The year the paper was published.
- **Study Period**: The average number of days the study was performed for each user. This was most often reported directly by the authors of the article themselves. When the period was not listed in days, we assumed that one week is seven days and one month is 30 days, to calculate an estimated number of days for the study. For example, "four sessions within two weeks" [86] has a study period of 14 days, while one session every week for a total of ten sessions [87] totaled a 70 day study period estimate (the sessions need not be on the same day per week). In rare cases where the study period was unclear, such as "for a semester," we listed the study period as unknown.
- **Number of User Sessions**: The average number of sessions between a robot and a participant, as relevant. Some studies were not session-based and instead were labeled as "free use" if use was entirely up to the user or "daily use" if participants were instructed to generally use the robot on a daily basis.
- **Session Length**: The average number of minutes per session (as relevant).

- **Total Interaction Time**: The average total time, in minutes, that each participant spent with the robot during the study. When this information was not given explicitly in a paper, we manually calculated an estimated time based on other characteristics listed in the paper.

Application Domains, Participants, & Study Location:

- **Domain**: The main application domain of the study. We classified studies into the following domains: physical health, mental and cognitive health, education, entertainment, service and workplace, general purpose (i.e., home robots that provide several general uses), and Autism Spectrum Disorder (ASD). Although ASD-related studies can overlap with physical or mental health domains, we categorized them separately due to the high volume of research specifically focused on ASD within the HRI literature.

- **Participants**: The number of data-generating participants in the study. We only consider the number of participants that interacted with a physical robot. For example, in [56], only 20 of 61 participants interacted with an embodied robot; the other participants interacted with systems that did not include a robot. Therefore, we considered the number of participants to be 20. We also do not consider participants who were excluded from the analysis for varying reasons, such as technical difficulties or leaving the study prematurely.

- **Age Group**: The primary age group of the participants. We categorized the participant samples into different age groups: infants (0–3 years), children (3–12), teens (13–17), adults (18–65), elderly (65+), and mixed. When more than 80% of the participants fell into one particular category, we classified it as the majority category instead of mixed.

- **Location**: The primary location where the study was conducted. We classified each study into the following locations: *home*, *school*, *care home*, *hospital*, *rehabilitation facility*, *daycare*, *laboratory*, *workplace*, *museum*, and *other*. We categorized a study under the *workplace* domain when the deployment took place in an office, business, or similar professional environment—for example, a corporate office for employees or a car rental agency. Although schools and hospitals are also workplaces for teachers and healthcare professionals, we classified these settings separately due to their unique social structures, institutional goals, and user populations. The nature of human-robot interaction in educational and clinical environments often differs substantially from that in general workplace settings, which warrants different categorization.

- **Country**: The primary country where the study occurred or participants were recruited.

Study and Robot Qualities:

- **Robot Platform**: The primary robot(s) or robotic platform(s) used in the study.

- **Autonomy**: The level of autonomy exhibited by the robot. We classified each system as *autonomous* (operating independently without human intervention), *semi-autonomous* (a combination of autonomous behavior and human control), or *non-autonomous* (fully operated by a human, such as through teleoperation or scripted control).

- **Interaction Dynamic**: The number and configuration of people involved in the interaction with the robot. We categorized interactions into six primary types: *dyadic* (one robot and one person), *triadic* (one robot and two people), *group* (one robot interacting with a group or family), *observer* (one person observing others interact with the robot; e.g., [88, 89]), and *mixed* (a combination of the above types within the same study; e.g., [54, 84]).

- **Personalization & Adaptation (Y/N)**: Whether the robot exhibited any form of personalization or adaptation. We labeled *Yes* if the robot adapted its behavior based on the user's actions or personalized its responses over time; otherwise, we labeled it as *No*.

Results & Measures:

- **Qualitative Results (Y/N)**: Whether the study had qualitative results. We labeled *Yes* if the paper presents a discussion or analysis of any qualitative findings, and *No* if the paper does not. The definition of qualitative for this review includes nonnumerical subjective measures, including but not limited to open survey responses, interviews, ethnographic methods, and behavioral observations. If some numerical data were collected but the predominant analysis, results, and discussion presented were qualitative, the paper was labeled qualitative only.

- **Quantitative Results (Y/N)**: Whether the study had results from a quantitative analysis presented. We labeled *Yes* if the paper presents a statistical analysis of quantitative findings, or *No* if the paper does not report any quantitative findings. At times, studies presented quantitative analyses on qualitative results. These were judged as quantitative in addition to qualitative if the results presented multiple numerical analyses towards the authors' claims. Studies presenting descriptive metrics (e.g., interaction counts or survey averages) alongside predominantly qualitative results were not considered quantitative. A small number of studies that collected longitudinal data for training machine learning models were also classified as quantitative.

- **Conditions (Y/N)**: Whether the study had conditions for the pursuit of statistically significant results. We labeled *Yes* if the paper presents a statistical analysis of separable and controlled conditions, or *No* if the paper does not have conditions or does not perform a statistical analysis of study conditions.

- **Long-Term Engagement (Y/N)**: Whether the study reported long-term engagement measures. We labeled *Yes* if the paper reports a measure of a user's engagement over time during the study, or *No* if not.

- **Engagement Measure**: The method(s) used to assess user engagement. For papers that reported engagement, we documented the specific measures used (e.g., self-reported surveys, study drop-out rates, frequency or duration of continued interaction with the robot).

2.4 Findings

This section presents an analysis of the 120 studies identified based on the inclusion criteria described previously. A complete list of these studies, along with a subset of key characteristics, is provided in Appendix A. As shown in Figure 2.2, the number of long-term HRI studies has increased steadily over the past two decades, accompanied by a broadening of the interaction types explored. In the following analysis, we examine the defining characteristics of long-term studies published since 2012, as introduced in Section 2.3, including study duration, participant demographics, interaction settings, modalities, and application domains. We also highlight approaches to measuring long-term engagement and strategies to improve outcomes, such as personalization and adaptation. Where relevant, we report trends across two time periods,

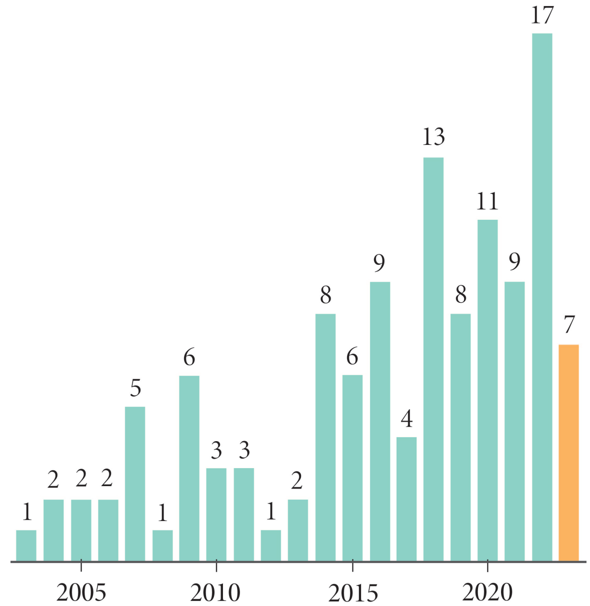


Figure 2.2: Annual Distribution of Reviewed Studies. The figure above shows the number of studies meeting our review criteria by year. The bar for 2023 reflects partial data collected between January and April, in contrast to the complete annual data available for previous years.

2003–2012 and 2013–2023, to illustrate how the field has evolved over time.

2.4.1 Temporal Qualities

In this section, we first analyze the various ways in which researchers have reported the temporal characteristics of long-term HRI studies.[3] These include the overall study period, the number of interactive sessions with a robot, the lengths of these sessions, and the total interaction time per user over the course of the study. As seen in Figure 2.3, these characteristics vary widely. For instance, some studies report on three sessions distributed over several weeks (e.g., [49, 91, 93]), while others (e.g., [5, 43]) investigate tens of sessions over multiple months.

Study Period

The *study period* is defined as the number of days from the first study session a participant has with the robot system to the final session with the system. Given the

[3]It is important to note that not all studies in our corpus make claims of being "long-term" despite meeting our qualifications for inclusion. Some authors utilize terms such as "multisession" (e.g., [90,91]) or "longitudinal" (e.g., [53,92]) to describe the nature of their study rather than "long-term." Our corpus includes all studies that met our criteria, separate from any authors' claims.

intentional breadth of our inclusion criteria for long-term interaction, we find a large variety of study periods ($M = 48.4$ days, $SD = 70.2$ days). These ranged from the minimum of three days [57] to 570 days [94], as illustrated in Figure 2.3a. For seven papers, which total eight studies, the study period was not reported [49, 85, 95–99].

Investigating the primary motivations of shorter versus longer studies, we observed several patterns. For studies shorter than two weeks ($N = 17$), we find that the objective of the study was often exploratory or in evaluating a new technique. For instance, nine of these studies aimed to investigate user perceptions and usability of a robot in a new environment (e.g., [1, 64, 100–102]) or the relationship built between participants and the robot (e.g., [51, 103]). Four of the studies that were shorter than two weeks sought to validate new technical methods (e.g., [57, 90, 104, 105]), and four studies show participant improvements in a new skill or task (e.g. [59, 100]). In contrast, many of the studies that occurred over the span of three months or longer ($N = 14$, [5, 6, 42–44, 56, 58, 85, 94, 106–110]) explored the adoption, engagement, and usage patterns of robots in long-term settings. This divide makes sense, as studies aimed at longer-term engagement and usage require that the study period last beyond any novelty effects.

By comparing study periods by decade, we find that more studies with shorter periods have occurred in the recent decade (2013–2023). For example, of the 23 studies in the first decade (2003–2012) that report a study period, only five (21.7%) have a study period of less than a month. In contrast, 48 (53.9%) of the 89 studies in 2013–2023 that report a study period report a period of less than a month. The increase in long-term studies over shorter periods of time can be considered in one of two manners: either the lengths of long-term studies have decreased over time, or there has been a natural progression from single-session studies into multisession studies. We find that there is an emergence of studies that do not intend to examine the long-term effects of the robot interaction, but rather to investigate certain elements of usability and instruction that require a handful of sessions instead of just one (e.g., [100, 101]).

User Sessions

In addition to the number of days of a study, we report on the duration and number of user sessions with a robot. These measures capture the amount of direct interaction that a participant has with a robot. A *session* is characterized by a specific duration when a participant is expected to interact with the robot system. 86 studies (71.7%) in

Figure 2.3: Comparison of Study Length and Frequency between 2003–2012 and 2013–2023. The distribution of the studies based on study length is shown in (a). For studies that were sessions-based rather than free use of the robot system, distributions of the studies based on the number of sessions (b) and session length in minutes (c) are shown. Lastly, the distribution of studies based on total study length in hours as reported or estimated by the reported number of sessions and session length is shown in (d). We compare the distributions across two decades to examine emerging trends in light of the rapid growth in long-term HRI research.

our corpus reported a particular number of sessions per participant and are therefore considered to be *session-based*. For sessions-based studies, we find that the number of sessions varies widely from three sessions [1, 49, 54, 93, 98, 111–113] to as many as 1559 sessions in the study by Ostrowski et al. [6], as shown in Figure 2.3b. To better represent the distribution of interaction frequency, we excluded the Ostrowski et al. study as a clear outlier. With this adjustment, the average number of sessions across session-based studies in our corpus is 15.5 ($SD = 23.9$). For 21 studies, the number of sessions per participant was not reported.

Among the session-based studies, 77 (representing 64.2% of the entire corpus) reported a specific session length. These durations varied considerably, ranging from as short as one minute [6, 103] to as long as three hours [85], with an average session length of 26.9 minutes ($SD = 28.2$ minutes). The remaining 32 studies did not provide a clear duration for each session. Consistent with our earlier findings on study periods (Section 2.4.1), we observe that the decade 2013–2023 saw a higher number of studies featuring shorter sessions and fewer total sessions compared to the previous decade (2003–2012), as illustrated in Figure 2.3c.

In our corpus, 10 studies were neither session-specified nor session-based, as they did not report the number or duration of the participants' sessions. An additional 11 studies reported that user sessions were self-directed *daily use* ($N = 9$, [110,114–121]) or *free use*[4] across the entire study period ($N = 2$, [41,102]).

Total Interaction Time

Three studies directly reported the total interaction time per participant [51, 84, 122]. Many of the remaining studies reported session lengths alongside the number of sessions. In order to compare sessions-based and free-use studies on the amount of participant-to-robot time, we present an estimate of the total interaction time per participant for each study. We estimated the total interaction time by multiplying the number of sessions by the session length for each study that reports both values ($N = 76$). Together, the 79 studies demonstrate an average interaction time per participant of 553.6 minutes ($SD = 1173.5$), ranging from 10 minutes [123] to 120 hours [43]. As seen in Figure 2.3d, we observe a slight decrease in the number of studies with fewer than two hours of interaction length published during 2013–2023 (N = 7, 41.2%) compared to 2003–2012 (N = 31, 50.0%).

2.4.2 Application Domains, Participants, and Locations

In this section, we classify our corpus of long-term HRI studies according to their primary domain, participant characteristics, and location of deployment (Figure 2.4). Early studies in long-term HRI focused mainly on educational settings [124,125], behavioral interventions for children with ASD [99,107], and general user perceptions of robots [84,89]. More recent work spans a wider range of applications—for example, using robots to support employees in office environments [88] or to promote cognitive health through physical activity [126]. As detailed below, we organize application domains into seven categories: education, mental and cognitive health, general purpose, physical health, entertainment and gameplay, and service and workplace.

Here, we also classify studies based on their participant age ranges (ranging from

[4]We define "free use" as users having complete control over when and how to use the robot. We acknowledge some researchers may not have reported casual guidance given to participants and that some participants may have felt implicit pressure to use the robot regularly.

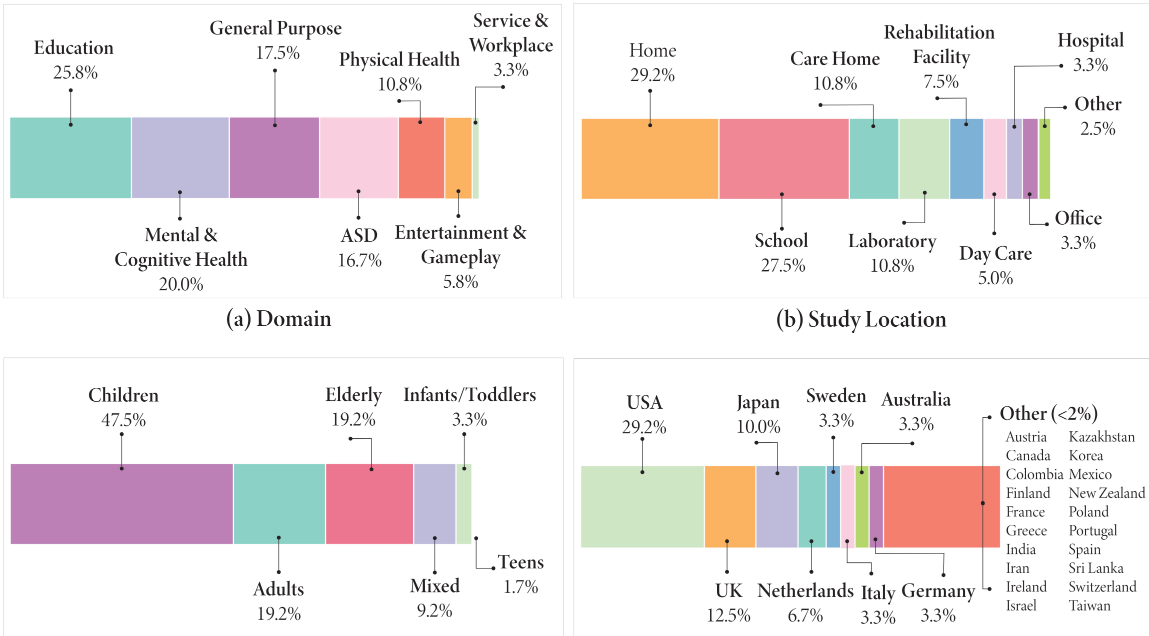
infants to the elderly) and the number of participants with reported data. Occasionally, the domain dictates the participant type of the study, such as children using an educational robot, though this is not always the case. Often, the location of a study is related to its domain and type of participants, such as a rehabilitation robot for the elderly that is deployed in a nursing home. Because long-term HRI research often seeks to simulate or prepare for use in the real world, understanding these contextual relationships is critical. Analyzing how the participant population, study setting, and application domain interrelate helps researchers evaluate not only the ecological validity of the deployment but also the scalability, relevance, and potential barriers to real-world integration. This contextual framing also helps to identify gaps, such as underrepresented populations or neglected environments, that may limit the generalizability or impact of current research.

Application Domains

We find that the long-term HRI studies in our corpus fall into the following domain classifications: *Education* ($N = 31$, 25.8%), *Mental & Cognitive Health* ($N = 24$, 20.0%), *General Purpose* ($N = 21$, 17.5%), *ASD* ($N = 20$, 16.7%), *Physical Health* ($N = 13$, 10.8%), *Entertainment & Game-Play* ($N = 7$, 5.8%), and *Service & Workplace* ($N = 4$, 3.3%). A summary of areas of study per domain is as follows:

- **Education ($N = 31$, 25.8%):** Within studies in the *Education* domain, robot interactions tend to be focused on tutoring or teaching skills in reading (e.g., [44,67,68]), math (e.g., [112,113,127]), language (e.g., [86,101,128]), handwriting (e.g., [129,130]), and other common academic subjects (e.g., [50,131]). For the majority of the studies in this domain ($N = 26$, 83.9%), the population of interest is children between the ages of 3 and 12, with the robot most often in a school ($N = 17$) or the home ($N = 6$).

- **Mental & Cognitive Health ($N = 24$, 20.0%)**: Studies within this domain can be organized into three categories: *elder care*, *condition-specific rehabilitation*, or *general wellness*. Ten studies target cognitive stimulation for elderly participants, ages 65 or older, either in the individual's home (N = 2; [100,117]) or care homes (N = 8; [60, 102, 132–136]). Condition-specific studies (N = 8; [43, 94, 101, 110, 115, 116, 137, 138]) examine the robot interactions with users with a diagnosed cognitive disorder such as Dementia, Down Syndrome, or Alzheimer's. The remaining seven studies [8, 53, 64, 103, 121, 126, 139] primarily examine improving mental and cognitive health in typically-developing adults.



(a) Domain

(b) Study Location

(c) Participant Age

(d) Country

Figure 2.4: Distribution of Study Qualities. These charts overview of the key attributes in long-term HRI studies conducted over the past two decades. The distribution is presented across four main dimensions: study domains (a), study locations (b), participant age groups (c), and countries where the studies were conducted (d). It is important to note that the largest category, "Other," in (d) encompasses 20 countries, each contributing less than 2% to our current dataset.

- **General Purpose ($N = 21$, 17.5%)**: Studies in the *General Purpose* domain are largely focused on investigating user perceptions of, and user engagement with, the robot. For instance, studies in this category have shown that users enjoyed the robotic interactions (e.g., [65, 96]), were motivated to interact with the robot (e.g., [97, 140, 141]), and demonstrated changes in their perception of (e.g., [1,85,142]) or engagement with the robot throughout the study (e.g., [6,42, 84, 92, 114]). Further studies examined methods for behavioral personalization and adaptation of the robot due to user preferences (e.g., [105]), used the robot as a tool for understanding the target population better (e.g., [85, 143]), or evaluated the feasibility of the robot to provide living assistance (e.g., [41]). *General Purpose* studies target a wide range of participant groups, from infants to seniors, in a variety of settings such as the home, a laboratory, daycare, or school.

- **ASD ($N = 20$, 16.7%)**: Studies in this category explore the use of robots with individuals with ASD. Several studies in the *ASD* category would traditionally fall within the *Mental & Cognitive Health* or *Education* categories; however, we treat these studies as a separate category due to their large number. Within studies in the *ASD* domain, most studies either investigate skills development or validation of new measurement or prediction techniques. Those in the skills group often investigate interactions such as verbal communication or social skills (e.g., [3, 5, 144, 145]). Studies about new measurement techniques focus on validating methods that predict engagement, build user models, or develop metrics for specifically the ASD population (e.g., [95, 99, 122, 146–148]). The majority of *ASD* studies focus on interactions with children ($N = 16$, 80.0%) in a school ($N = 6$) or home ($N = 5$) setting. Within the past half-decade specifically (2018–2023), only a small number of studies have extended these efforts to adolescents [5, 149] or adults with ASD [59], highlighting a significant gap in current research. Although robots have garnered significant attention for use in ASD interventions, most existing work remain focused on younger children and are not designed to support the evolving social needs of individuals with ASD across the lifespan.

- **Physical Health ($N = 13$, 10.8%)**: Studies within the *Physical Health* domain primarily fall into two categories: *general wellness* and *condition-specific rehabilitation*. For general wellness, studies focused on supporting exercising and healthy eating ($N = 8$, e.g., [52, 150–152]) across a wide range of participant age groups, but primarily centered on adults and older populations ($N = 5$). These robots are often deployed in a home or laboratory setting. Condition-specific studies ($N = 5$) instead support individuals recovering from a medical event such as surgery, stroke, or a diagnosis of a life-impeding condition [57, 58, 153–155]. Robots for condition-specific support are more often found in hospitals or rehabilitation centers, almost entirely for elderly users ($N = 4$).

- **Entertainment & Game-Play ($N = 7$, 5.8%)**: The studies in this domain [2, 66, 90, 108, 109, 156, 157] primarily investigate the potential for engaging and maintaining user interest with a robotic system via games or other interactive media. A majority of studies in this domain ($N = 5$) focus on children younger than 12 years old in a range of settings, such as a laboratory, school, daycare, or home.

- **Service & Workplace ($N = 4$, 3.3%)**: In our corpus, there are currently only four studies [84, 88, 123, 139] that investigate robots in the *Service & Workplace* domain. These robots are designed to perform tasks related to workplace offices or other locations that offer services to customers. For example, Vishwanath et. al. [88] explored how a humanoid robot receptionist could improve staff productivity in an office setting. While many studies included in the *Education*, *Physical Health*, and *Cognitive Health* domains are conducted in settings that arguably offer services (e.g. schools and health clinics), these studies focus on a specific educational or health outcomes rather than on the workplace itself.[5]

Upon evaluating study domains over the past 20 years, we find several trends over time. While the first five years of long-term studies (2003–2007) consisted of a small number of studies ($N = 12$), these studies were evenly distributed across six domains: *Mental & Cognitive Health*, *ASD*, *General Purpose*, *Entertainment & Game-Play*, *Education*, and *Service & Workplace*. In contrast, 50% of the studies published in the following five years (2008–2013) were in the *General Purpose* domain. A newfound interest in the *Physical Health* domain emerged in 2009 [52] and has steadily grown to the present year. *Physical Health*, *Mental & Cognitive Health*, *Education*, and *ASD* all have exhibited steady growth over the past fifteen years, whereas the *Service & Workplace* domain has only recently emerged for long-term study.

Many of these domains are inherently suited for long-term study. For instance, with *Education*, *ASD*, and to a certain extent *Mental & Cognitive Health* and *Physical Health* studies, often the goal of the robotic intervention is to support the acquisition of skills. Developing a new skill, by its nature, requires time to develop and must be studied as such to show true value. Similarly to skill acquisition, therapies such as rehabilitation and cognitive exercises also require repeated practice of a certain activity. In contrast, *General Purpose* studies are often more related to the exploration of how social robots may integrate into our homes and daily lives. This type of study requires longitudinal study, as real-world robot applications will not happen in a single session. For researchers interested in understanding the adoption of robots in society, long-term study is an important step. In the less represented domains of *Entertainment & Game Play* and *Service & Workplace*, such real-world requirements still apply, but there may be other challenges. As the *Service & Workplace* domain is still relatively new in long-term exploration, researchers may opt to first explore the

[5] For example, we classify work such as Rueben et. al. [123] within the *Service & Workplace* domain because the authors investigated the impact of a mobile shoe rack at a yoga studio on client satisfaction. Had the authors investigated the impact of the robot on the individual performance of yoga students at the studio, we would have classified this work under *Physical Health*.

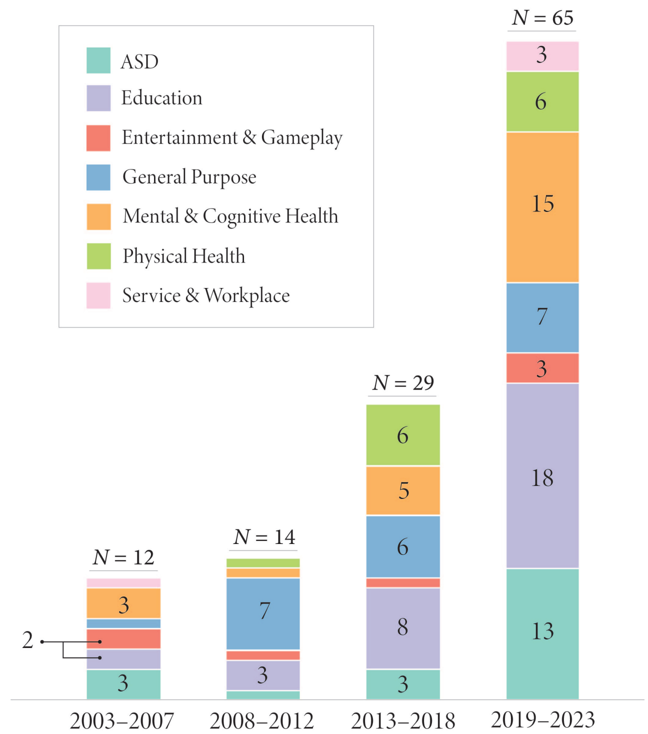


Figure 2.5: Distribution of Application Domains. The chart illustrates changes in the distribution of studies across application domains over time, grouped in five-year intervals.

problem space, initial usability, and initial user perception with a single-session study before transitioning to long-term study. The *Entertainment & Game Play* domain offers an easy way to test user perceptions of a robot and algorithmic contributions around engagement; however, researchers may have less interest in this domain now as more specific and beneficial use cases for social robots have emerged (e.g. supporting social skills development, rehabilitation, home assistive tasks, etc.).

Participant Types

For the purposes of this review, we classify the participants according to their age group corresponding to the general stages of psychological development ranging from infants to seniors. Across our corpus, 57 studies (47.5%) feature children (ages 3 to 12) as the primary participant type, 23 (19.2%) feature adults (ages 18–65), 23 (19.2%) feature seniors (above age 65), four (3.3%, [108,137,140,158]) feature infants or toddlers (under age 3), and two studies (1.7%, [5,149]) feature teenagers (ages 13 to 18). The remaining 11 studies (9.2%, [42,43,52,53,67,86,96,97,104,123,153]) report outcomes of the robot interaction to describe mixed demographics across participant age groups.

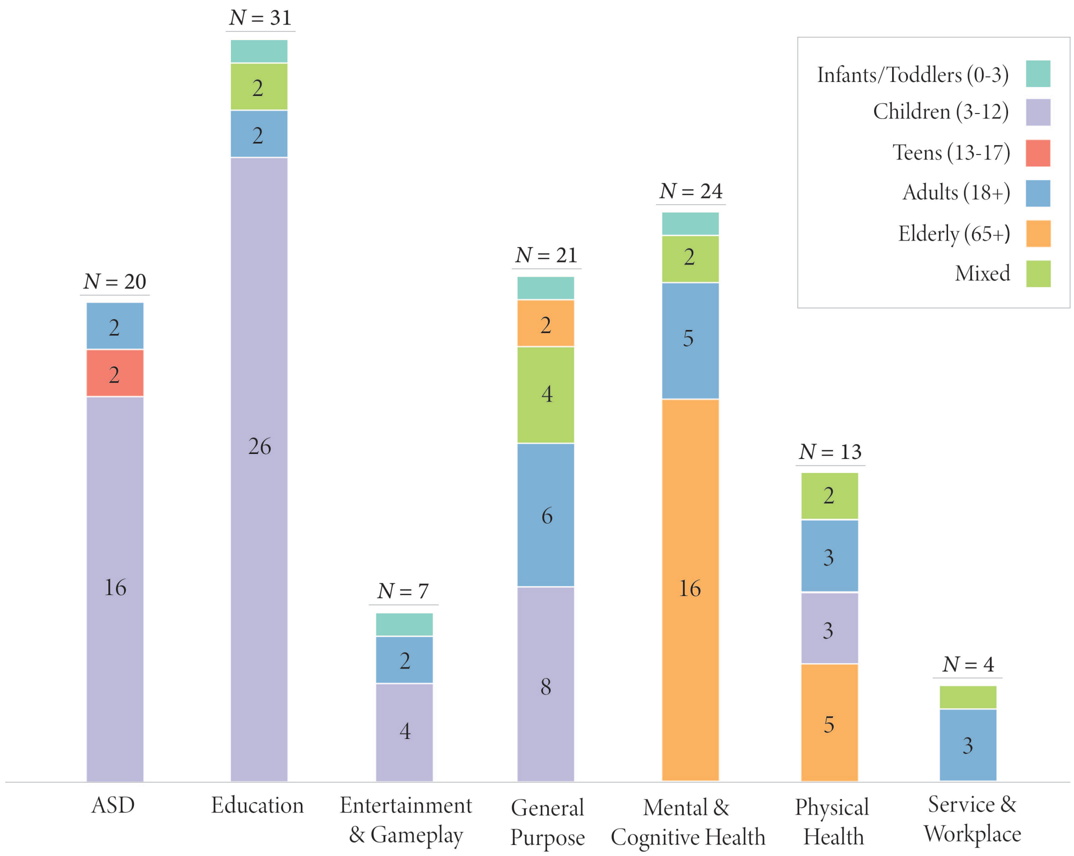


Figure 2.6: Participant Age Distributions Across Domains. This chart illustrates how participant age groups are distributed across domains. Study domains often reflect the population of interest, and vice versa.

It is expected that the domain of a study is related to the population of interest of the study, or vice versa (Figure 2.6). For example, the majority of *Mental & Cognitive Health* studies target the senior population ($N = 16$, 64.0%) because many studies in this domain target cognitive stimulation for users with a diagnosed cognitive disorder such as Dementia or Alzheimer's. Similarly, the vast majority of studies within the *Education* and *ASD* domains focus on children ($N = 42$, 82.4%). Studies in the *General Purpose* and *Physical Health* domains contain a diverse range of participant age groups.

Participant Counts

In our corpus, the number of participants per study varies widely ($M = 24.8$, $SD = 30.4$), from a single participant [58, 84, 100, 137] to as many as 228 participants [124]. The distributions of participant counts differ by domain: *ASD* ($M = 14.8$, $SD = 13.0$), *Education* ($M = 36.8$, $SD = 44.4$), *Entertainment & Game Play* ($M = 15.3$,

SD = 10.5), *General Purpose* (M = 35.8, SD = 32.8), *Mental & Cognitive Health* (M = 19.5, SD = 21.9), *Physical Health* (M = 12.8, SD = 9.4), and *Service & Workplace* (M = 11.3, SD = 10.8). The majority of the 14 studies with less than five participants [5, 58, 84, 85, 93, 100, 103, 104, 107, 137, 144, 147, 153, 159] were focused on providing therapies to protected or sensitive populations, such as those with diagnosed physical or mental disabilities (N = 10). On the other hand, for the six studies with more than 100 participants [7, 96–98, 124, 160], research was carried out through educational systems [98, 124, 160] or with an off-the-shelf commercial home robot [96, 97]. The four studies with the largest sample of participants [7, 98, 124, 160] were conducted in an educational setting and for children. Although these studies are outliers in their respective domains due to their sample size, conducting studies through established school structures likely provides easier access to large numbers of student participants. Similarly, using a preexisting commercial home robot with minimal customization likely reduces the burden of designing and deploying systems in the unstructured and diverse environment that is the users' homes.

The number of participants directly impacts the types of statistical analysis that can be performed, as many methods require a minimum sample size to produce reliable and valid results. Among studies reporting statistical results (N = 70; e.g., between experimental conditions or between different population segments), the distribution of participant sample sizes tends to larger counts (M = 31.7, SD = 35.6), with 65.7% with 15 participants or more. Among the remaining studies that do not conduct statistical analyses (N = 50), the distribution of participant counts tends toward smaller counts (M = 17.0, SD = 19.3), with 72.0% below 15 participants. The motivation, expected outcomes, and structure of a specific study can inform the analyses (e.g., qualitative, quantitative or statistical) researchers choose to conduct. In Section 2.4.4, we examine the potential factors that influence the results reported in our corpus of studies.

Study Locations

As long-term studies seek to emulate more real-world scenarios for HRI, the setting where robot interactions occur is an important study design consideration. Studies performed in a research facility or laboratory have the advantage of controlling for environmental variables in order to isolate specific components of interaction. In contrast, "in-the-wild" environments such as homes or classrooms are more likely to produce findings that are generalizable to real-world interactions and contexts.

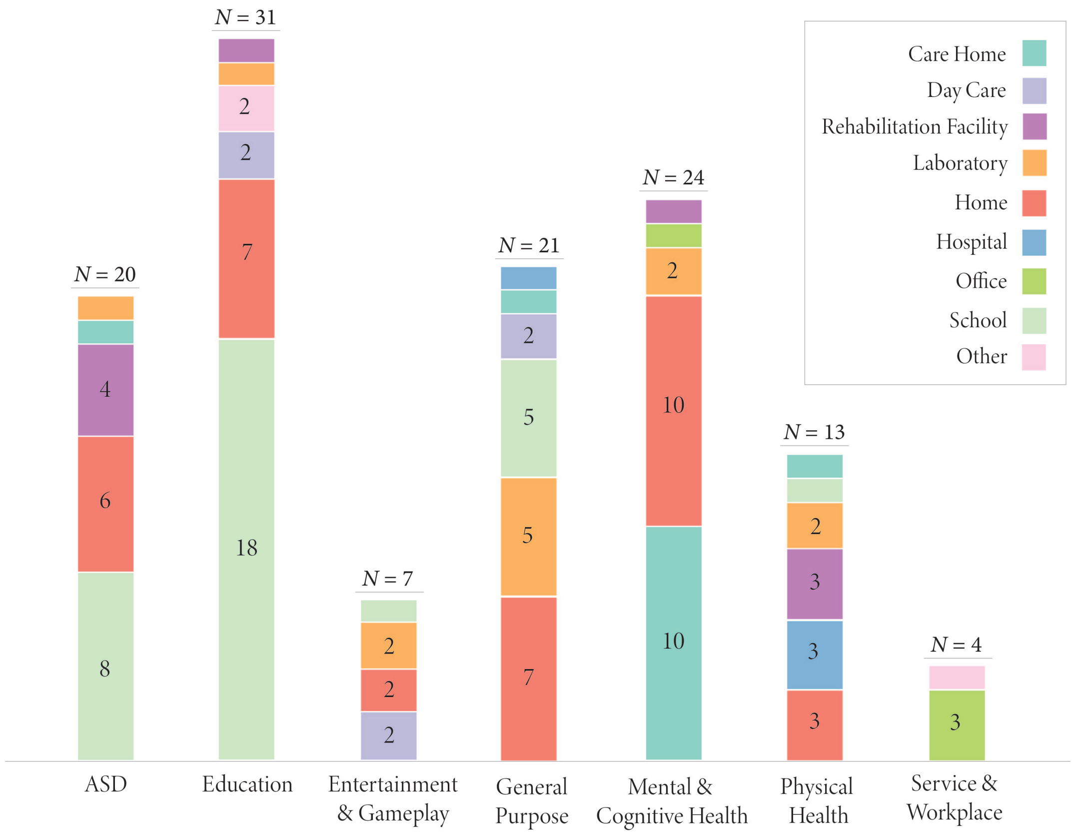


Figure 2.7: Study Locations Across Domains. The distribution of study location by domain is illustrated. The location of a study typically aligns with the study's domain.

However, in-the-wild environments are dynamic and unstructured settings and thus present greater technical demands and challenges for robot use, design, and deployment.

In our corpus, most of the studies ($N = 103$, 85.8%) were conducted in real world settings, most commonly in participants' homes ($N = 35$, 34.0%) or in schools ($N = 33$, 32.0%). The remaining 17 studies (14.2%) took place in laboratory environments. Of these, 10 studies [1, 53, 85, 86, 95, 104, 137, 151, 157, 161] were conducted in lab settings specifically designed to simulate naturalistic environments, such as homes or workplaces. A complete breakdown of study locations is shown in Figure 2.4b.

As with participant types, the location of a study usually aligns with the study domain, as shown in Figure 2.7. For instance, many educational robots are deployed in schools (e.g., [48, 50, 130, 162–165]), with a subset deployed in homes for tutoring outside of the classroom (e.g., [44, 120, 166, 167]). Similarly, robot interactions that target physical health rehabilitation are often conducted in hospitals or care facilities (e.g., [51, 58, 153]).

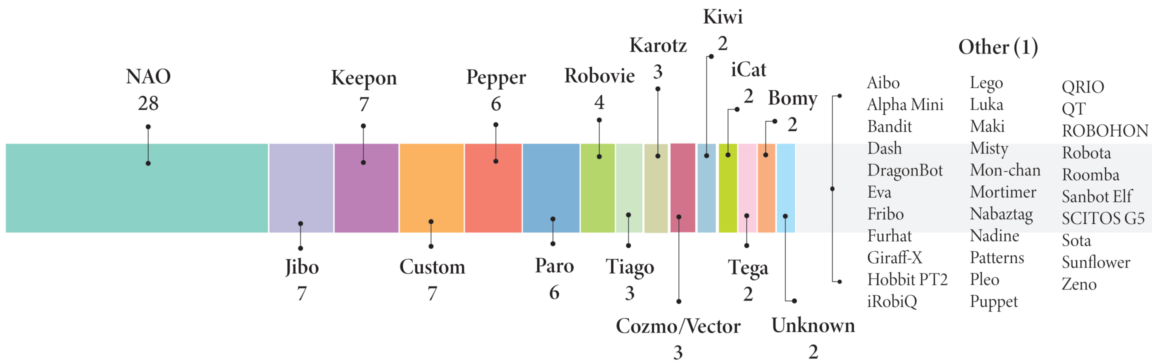


Figure 2.8: Robot Platforms. Illustrated is the distribution of robots employed in long-term human-robot interaction studies. The largest category, "Other," encompasses 34 platforms that were each represented only once in our dataset.

Beyond the immediate environment of the robot, the geographic and cultural environment can influence the results of long-term studies. In our corpus, we find that 41.7% ($N = 50$) of the studies were conducted in Europe, mainly in the United Kingdom ($N = 15$); 32.5% ($N = 39$) of the studies were conducted in North America, primarily in the United States ($N = 35$); 18.3% ($N = 22$) of the studies were conducted in Asian countries, mainly in Japan ($N = 12$); and 5.0% of the studies were conducted in Australia ($N = 4$) or New Zealand ($N = 2$). The remaining studies do not report the location [4, 88, 95].

2.4.3 Study and Robot Qualities

Researchers have explored a wide range of long-term human-robot interaction modalities across different domains, participant populations, and study settings. In this section, we identify patterns in our corpus related to the robot platform used, its level of autonomy, the interaction dynamic (e.g., dyadic, triadic, group), and whether the robot employed behavioral adaptation or personalization. By categorizing these core dimensions of long-term HRI, we reveal the diversity of interaction formats and shed light on how long-term engagements have traditionally been implemented and studied.

Robot Platform

A robot's physical form and affordances play a crucial role in shaping the types of interactions that can be studied. Using commercially available (i.e., "off-the-shelf") robots offers several advantages, including physical durability, ease of deployment, and the ability to compare findings across studies that use the same platform. However, these robots often come with inherent design constraints that can limit their flexibility or suitability for specific research objectives. For instance, Ramnauth et al. [59] supplemented the widely used Jibo robot [168] with additional sensors to compensate for inaccessible hardware features, highlighting how even popular platforms may require adjustment to meet the objectives of a study.[6]

Upon investigation, we find that the NAO robot [169] is the most commonly used platform in long-term HRI studies (e.g., [50, 91, 129, 146, 147, 151, 163]), with 30 (25.0%) studies reflecting its widespread availability, ease of programming, and suitability for a variety of interaction contexts and user populations. The next most popular robots include: Keepon ($N = 7$, 5.8%; [54, 85, 111, 126, 162]), Jibo ($N = 7$, 5.8%; [3, 6, 8, 59, 64, 67, 166]), Paro ($N = 7$, 5.8%; [60, 94, 101, 106, 132, 133, 138], and Pepper ($N = 6$, 5.0%; [49,51,53,114,118,153]. We provide a brief description of these most popular robots below. Beyond these five commercially available platforms, only seven studies (5.8%; [52, 84, 92, 121, 123, 156, 158]) feature novel, custom robots or prototypes.

- **NAO** (25.0%; $N = 30$): NAO [169] is a small, programmable, table-top humanoid robot with a rich sensor suite and built-in interactions for complex natural language and facial and gesture recognition. Its popularity in HRI research is likely due to its diversity of affordances and capabilities, the ability to easily purchase the robot off-the-shelf, as well as its existing popularity lending itself to easier research reproducibility.

- **Keepon** (5.8%;$N = 7$): Keepon [170] is a small, minimalist tabletop robot with a soft, expressive body capable of moving side to side, up and down, and rotationally. Although it features audio and visual input and is relatively affordable compared to more complex robots, the commercially available Keepon was a passive device without onboard sensing or computation. Unlike platforms such as Nao or Jibo, it was not a standalone, off-the-shelf system and could not be directly used in HRI studies without significant external augmentation.

- **Jibo** (5.8%; N = 7): Jibo [171] is a tabletop robot with a more cylindrical form-factor and a circular display that can rotate expressively in multiple directions. Similar to NAO, it is equipped with audio and video inputs for complex natural language and computer vision.

[6] This work is presented as Chapter 6. Additional hardware adjustments to the Jibo platform are also featured in Chapters 5 and 9.

- **Paro** (5.8%; N = 7): Paro [172] is a soft and plush robot designed to look like a seal and responds to touch, light, and sound in order to express certain emotional states.
- **Pepper** (5%; N = 6): Pepper [173], a semi-humanoid on wheels, is the largest form factor of the top four robots, reaching almost human height with the ability to navigate an environment. It features a touchscreen on its chest, multimodal sensors (cameras, microphones, depth sensors), and gestural capabilities. While offering a rich interaction interface, Pepper's resultantly high cost and mechanical complexity can pose limiting factors.

Robot Autonomy

Future real-world robots intended for long-term deployment in homes, workplaces, and other everyday settings will need to operate with full autonomy. However, achieving robust autonomy in natural, uncontrolled environments presents significant technical and perceptual challenges. Consequently, the decision of whether—and to what extent—a robot should be autonomous is a critical design consideration for researchers.

In our corpus, we observe the use of non-autonomous, semi-autonomous, and fully autonomous systems. While a *fully autonomous* robot is most aligned with real-world applications where the robot's architecture alone solely directs its behaviors, researchers may choose to implement a *non-autonomous* system via Wizard of Oz[7] or teleoperated techniques [174]. Such techniques enable researchers to imply complex levels of autonomy without the technical requirements of building an autonomous system. In between non-autonomous and fully autonomous systems lies *semi-autonomous* systems, in which researchers have access to a human-in-the-loop method of updating robot behaviors while deployed. Such a system can enable researchers to correct robot errors or introduce deeper levels of personalized interactions.

We find that 91 studies (75.8%) in our corpus utilize *fully autonomous* systems, 18 (15.0%) use *non-autonomous* systems, and 10 (8.3%) use *semi-autonomous* systems. We further observe a remarkable growth of 450% in autonomous design between 2003–2012 (N = 14) and 2013–2023 (N = 77). We do not observe this significant growth in semi-autonomous or non-autonomous design between 2003–2012 ($N = 12$) and 2013–2023 ($N = 16$).

[7] A method in which participants interact with a robot system that users believe to be autonomous, but is actually operated or partially operated by another human

Such growth may reflect the field's increasing desire to explore the open questions and practical challenges of deploying interactive robots in real-world, long-term settings. A central issue in these contexts is sustaining user engagement without the presence of a researcher to guide or scaffold the interaction. As discussed in Section 2.4.3, designing effective autonomous behaviors depends heavily on the nature of the robotic system—its physical design, sensor suite, interaction modalities, and computational capabilities. A simpler robot may be easier to deploy and more robust in uncontrolled environments but may offer a limited behavioral repertoire, potentially reducing its capacity to maintain meaningful or varied interactions over time. In contrast, a more complex robot can support a broader range of behaviors and interactions, but introduces challenges related to usability, user comprehension, and content design. If the robot's functionality is too opaque or overwhelming, users may struggle to engage effectively.

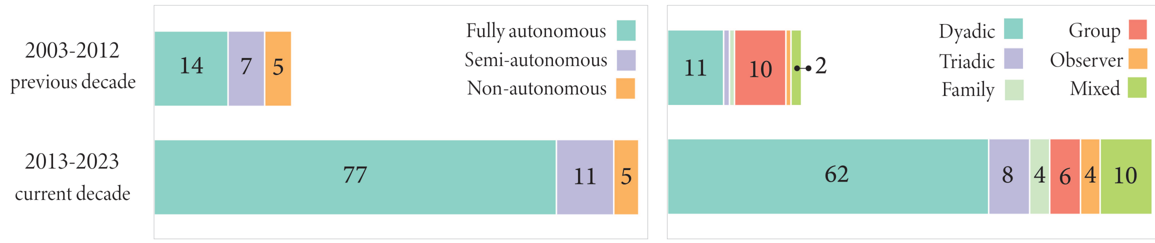
Another fundamental design decision in the development of autonomous behavior concerns how *proactive* the robot should be. A robot that is too passive may fade into the background and be ignored, while a robot that initiates too frequently or at inopportune moments risks becoming intrusive, irritating, or socially inappropriate. Determining the right balance—where a robot can recognize opportunities for meaningful engagement and respond appropriately—is an ongoing challenge and a rich area for research.

These complexities illustrate that designing autonomous robot behavior is not only a technical endeavor but also one that involves many social and psychological considerations. As such, the development of autonomous behaviors that are context-sensitive, adaptive, and user-aware remains a crucial and fertile frontier for long-term HRI research.

Interaction Dynamic

Human-robot interaction can take on many forms with respect to the social configuration of the interaction. We investigate the differences in studies that are *dyadic* (one-on-one between the human and robot), *triadic* (two humans interacting with one robot), *family* (one robot in a household interacting with multiple family members), *group* (one robot interacting with multiple humans in a group setting), *observer* (the human is observing a different human or group interacting with a robot) and *mixed* (a combination of any of the prior four categories) interactions.

In general, most long-term studies are dyadic ($N = 73$, 60.8%), with 16 studies



(a) Robot Operation by Decade

(b) Interaction Type by Decade

Figure 2.9: Robot Operation and Interaction Types. The charts display the distributions of robot operation (a) and study interaction dynamic (b) represented in our corpus, organized by decade.

(13.3%) of group interactions, 12 studies (10.0%) of mixed interaction dynamics, nine studies (7.5%) of triadic interactions, five studies (4.2%) of family interactions, and five studies (4.2%) of observer interactions. In comparing the distribution of interaction dynamics between the previous decade and 2013–2023 (Figure 9.4), we see a shift in the percentage of dyadic studies and group studies. Dyadic studies have grown in representation, from 42.3% to 66.0%, while group interaction studies have decreased in representation, from 38.5% to 6.4%.

One might have anticipated that as the field of long-term HRI matured, it would naturally expand toward more diverse interaction dynamics beyond dyadic (one robot, one user) configurations. However, our corpus suggests that dyadic interactions remain dominant. Several plausible factors may contribute to this trend.

First, longitudinal studies are inherently resource-intensive and, therefore, designing for a single-user interaction significantly reduces complexity. With dyadic studies, researchers can tailor the robot's behavior, dialogue, and sensing to a single participant without having to account for the added variability of group dynamics, such as turn-taking, social hierarchies, or shifting roles. This simplification extends to logistical and ethical considerations as well—recruiting and obtaining consent from one participant is much easier than from multiple group members, especially in sensitive domains.

Second, the prevalence of studies in the education and health domains in the past decade may help explain the persistence of dyadic formats. These domains often center on individual outcomes, such as a student's learning gains or a patient's behavioral improvements. While caregivers, teachers, or therapists are frequently involved in the broader intervention, the primary outcome measures tend to focus on the performance or behavior of a single individual, making a dyadic setup more practical for both design and evaluation. In contrast, general-purpose robots intended for home use are more likely to interact with entire families, and workplace robots may need to support multi-user coordination or collaboration, requiring more sophisticated interaction dynamics.

Finally, the rise of quantitative research methodologies over time may also play a role, as we first mentioned in Section 2.4.4. Quantitative metrics, such as gaze duration, task performance, or usage logs, are easier to isolate and interpret in single-user contexts, where tracking and attribution are straightforward. Likewise, personalization and adaptation techniques are more feasible to implement and maintain over time when tailored to an individual's unique profile, preferences, or developmental trajectory in dyadic interactions rather than to that of a fluctuating group of users. We discuss this further in Section 2.4.3.

Taken together, these factors suggest that the prevalence of long-term dyadic HRI studies reflects not just historical precedent but also practical, methodological, and domain-specific considerations that continue to shape how interactions are designed and studied.

Personalization and Adaptation

Personalization [175] and adaptation [176], two methods to tailor the robot's behaviors and interaction with a particular user, have been shown to improve outcomes for long-term HRI [141, 162, 177]. Such methods foster a sense of rapport, trust, and enjoyment in users, leading to more meaningful and engaging interactions. In addition, personalization and adaptation can allow the robot to maintain its relevance and usefulness over time, adapting to varying contexts, user preferences, and situational demands. We find that 47 studies (39.2%) in our corpus utilize personalization or adaptation in their robot design. They are roughly evenly distributed across domains.

Some common methods for personalization and adaptation include adjusting task difficulty based on prior performance (e.g., [3, 48, 127]) or utilizing affect recognition or physiological data to improve user engagement (e.g., [90, 135]). Many studies specifically investigate the effects of personalization and adaptation methods as their main research objectives (e.g. [48,49,68,98,105,112,113,141,157,159,162,177]). These studies have demonstrated that adaptive or personalized system results in increased user satisfaction, improved performance and effectiveness, enhanced user engagement, and improved flexibility in changing environments.

Overall, we observe a marked increase in the proportion of long-term HRI studies that incorporate personalization and adaptation techniques. In our corpus, 43% ($N =$

40) of studies from the most recent decade (2013–2023) include some form of user-specific behavior, compared to only 23% ($N = 6$) in the preceding decade (2003–2012). Several factors likely contributed to this upward trend.

First, as long-term HRI research has matured, studies have begun to shift from exploratory proof-of-concept deployments to more targeted investigations of specific interaction strategies. With engagement generally expected to decline over time—especially in real-world, unsupervised settings—there is growing recognition that maintaining user interest requires systems that can adapt meaningfully to individual users. Personalization offers a promising strategy to mitigate habituation and improve the perceived relevance and effectiveness of robot behaviors in repeated interactions.

Second, the technological landscape has evolved considerably over the past two decades. Advances in computer vision, natural language processing, biometric sensing, and real-time data analytics have made it significantly easier to gather and interpret user-specific information. These tools enable more sophisticated user modeling and support dynamic behavioral adjustments based on user preferences, affective state, skill level, or prior interactions. The increasing accessibility of these technologies reduces the barrier to entry for the implementation of adaptive systems in HRI research.

Finally, the growing interest in user-centered and inclusive design principles within the HRI community may also play a role. Personalization aligns well with the larger goals of creating socially intelligent systems that can recognize and respond to diverse user needs, backgrounds, and abilities. As a result, adaptive behaviors are no longer viewed as experimental extras but increasingly as core components of a successful long-term interaction design.

2.4.4 Result Types and Measures

Long-term HRI studies are varied in their type of research findings, with a mixture of quantitative and qualitative insights. In our corpus, case studies and ethnographies are present alongside quantitative studies and those that seek statistically significant differences between conditions. In this section, we outline the ways long-term HRI researchers have pursued different types of measured results. We additionally report patterns on two specific measures that are particularly relevant to long-term studies: pre/post analyses and long-term engagement.

Qualitative versus Quantitative Approaches

A fundamental decision long-term HRI researchers must make is whether to pursue a qualitative or quantitative methodology, or a blend of both approaches. This choice is multifaceted and hinges on several motivations that must be carefully weighed. Qualitative research often involves exploratory research questions rather than setting experimental conditions to test a hypothesis. Reporting qualitative data can provide insights into user preferences, challenges, and needs as well as serve as a foundation for new hypotheses or theories [59,123,156]. Examples of qualitative measures in our corpus include interviews, open-ended questionnaires, experimenter observation, video labeling, and diaries. In contrast, quantitative research typically facilitates objective measurements and evaluations of specific outcomes or performance metrics. Examples of quantitative measures in our corpus include: test performance, game scores, robot usage rates, interaction type counts, standardized surveys, custom surveys often containing Likert or similar scales, etc. Such metrics can be crucial in assessing the effectiveness or efficiency of robot interventions, measuring user perceptions, or evaluating task completion rates.

In our corpus, 49 (40.8%) studies report both qualitative and quantitative results, 33 (27.5%) report only qualitative results, and 25 (20.8%) studies report only quantitative results. The measures and metrics employed by researchers are often specific to the domain, participant type, and setting, and we encourage readers to consult Appendix A to find long-term studies with similar deployments as examples.

We find that the percentage of studies with quantitative results (either quantitative only or both quantitative and qualitative) has risen over the years, with 58.3% ($N = 70$) in the current decade versus 13.3% ($N = 16$) in the prior decade. With regard to domains with at least 10 total studies: we find that 77.4% of *Education* studies, 76.9% of *Physical Health* studies, 75.0% of *ASD* studies, 72% of *Mental & Cognitive Health* studies, and 68.2% of *General Purpose* studies have quantitative results. In contrast, the percentage of studies with qualitative results (either qualitative only or both quantitative and qualitative) has decreased from the previous decade (76.9%, $N = 20$) to the current decade (67.0%, $N = 63$). We find that the distribution of these results types varies more widely across domains in comparison to quantitative results. The large majority of *Mental and Cognitive Health* (84.0%, $N = 21$), *Physical Health* (76.9%, $N = 10$), and *ASD* (75.0% $N = 15$) studies contain qualitative studies, whereas a smaller majority or minority of *General Purpose* (63.6%, $N = 14$) and *Education* studies (48.4%, $N = 15$) contain qualitative results.

These differences may be explained by the inherent nature of the application domain. For instance, the *Education* domain often features clearly defined learning goals, student testing protocols, and standardized measures of academic achievement, making it more amenable to quantitative evaluation. Additionally, education-focused HRI has a longer history and is likely moving from exploratory system-building toward more rigorous assessments of learning outcomes and efficacy, thereby favoring more quantitative metrics. In contrast, *Mental & Cognitive Health* is a relatively newer domain in HRI, particularly for long-term interactions. Many of these studies focus on underserved or vulnerable populations, including individuals experiencing stress, cognitive decline, or social isolation. These contexts often benefit from early-stage qualitative research to understand complex behavioral changes, subjective well-being, and contextual factors that are not easily reduced to numbers. Researchers may prioritize narrative accounts, interviews, and observational data to assess therapeutic relevance or emotional resonance before scaling to larger quantifiable trials.

The *ASD* domain sits at a unique intersection of both approaches. On the one hand, the field benefits from well-established clinical benchmarks and standardized diagnostic instruments (e.g., ADOS, Vineland, SRS), which facilitate robust quantitative assessment of social, cognitive, and behavioral outcomes. On the other hand, individuals with ASD exhibit high heterogeneity in abilities, needs, and preferences, which complicates broad generalization and demands fine-grained, individualized interpretation. Moreover, many interventions rely on caregiver, teacher, or therapist reports to contextualize the child's behaviors—often requiring rich qualitative insight. As a result, HRI research in the *ASD* domain frequently combines both qualitative and quantitative measures. For example, gaze tracking, turn-taking frequency, or response latency can be paired with parental interviews, annotated video logs, or open-ended caregiver feedback. This mixed methods approach allows researchers to assess not only *what* changed over time, but also *why* the intervention may have succeeded or failed for a particular individual. The use of both types of data is especially crucial for capturing the nuances of long-term change, emotional trust, or developmental shifts that may emerge subtly and gradually.

Given the increasing interest in deploying robots in homes and clinics for autism therapy, this domain is particularly well positioned to benefit from nuanced evaluation frameworks that recognize both standardization and individual difference. As the field evolves, future studies may further explore hybrid methodologies that incorporate adaptive personalization with both subjective and objective evaluation tools to better capture the complex trajectories of individuals on the spectrum.

Study Conditions

For quantitative studies, conducting experimental conditions can help establish causal relationships between variables and make comparisons between different conditions. By including a control condition, researchers can determine whether the experimental manipulation or intervention likely causes the observed effects. In our corpus, 47 studies (39.2%) used quantitative methods with preset experimental conditions in pursuit of statistical significance. Often, these conditions were on (a) the use of a robot versus no robot (e.g., using a robot with a set of smart sensors for elder care versus just the sensors themselves [117]), (b) the inclusion of a specific interaction or approach versus without (e.g., including personalization or not with a general purpose robot for children [141]), or (c) between different types of populations (e.g., special needs children and typically developing children [55]). We find that the percentage of studies with conditions-based experimentation has increased in the current decade (43.6%; $N = 41$) compared to the previous (23.1%; $N = 6$), likely following the trend of increasing quantitative studies.

Pre/Post Experiment Analyses

To evaluate the long-term effects of robotic interventions, a common approach is to analyze specific metrics collected in the same way before and after the experiment. Among the studies in our corpus that employed quantitative analysis, 20.9% ($N = 18$) utilized such pre/post experiment comparisons.

There are multiple common methods used for this approach. One such method is Applied Behavior Analysis (ABA), commonly employed in clinical and psychological research to evaluate how specific interventions influence behavioral outcomes. For example, Jeong et al. [8] used ABA to demonstrate that a companion-like robot significantly improved participants' psychological well-being, while Scassellati et al. [3] found that robot-assisted interventions improved related clinical scores in children with ASD. In educational contexts, the most common metric involves comparing educational test responses and scores before and after robotic deployment (also known as "pre-tests" and "post-tests"). For instance, several studies assessed the effectiveness of fixed versus personalized tutoring assistance [48, 113, 162, 164] and examined the impact of different scaffolding behaviors exhibited by robots [163]. Beyond formal testing, other approaches to before-and-after assessment include measuring behavioral changes in school-age children over the course of interaction sessions [118], analyzing shifts in attitudes or perceptions through pre- and post-intervention questionnaires [1,139], and collecting interview-based feedback to gain qualitative insights into participants' experiences and behavioral developments [120].

Measuring Long-Term Engagement

While long-term HRI studies employ a wide range of qualitative and quantitative measures, we highlight one particularly salient metric in this review: long-term engagement (LTE). In our corpus, 45 studies (37.5%) explicitly measured LTE in some form. This metric has particular relevance for long-term HRI research for several key reasons.

First, LTE serves as a critical indicator for determining whether a study has moved beyond the novelty effect, as described in Section 2.2.1. By tracking changes in user engagement over time, researchers can assess whether observed outcomes are sustained or are merely artifacts of initial user interest. Without accounting for this temporal factor, studies risk misattributing early positive responses to the robot's design or effectiveness, rather than to transient novelty.

Importantly, six studies in our corpus explicitly attempted to measure or characterize the role of novelty in shaping participant behavior or outcomes [41–43, 53, 100, 115]. These efforts underscore the importance of engagement as both a research outcome and a methodological checkpoint in longitudinal work.

LTE is also central to the design and evaluation of adaptive and personalized systems (as detailed in Section 2.4.3). By capturing longitudinal engagement patterns, researchers can gain deeper insight into which interaction strategies sustain user interest and satisfaction over time—insights that are critical for refining system behavior and enhancing the overall quality and relevance of the HRI experience. In our corpus, 24 studies explicitly link personalization or adaptation with engagement outcomes, demonstrating how user-tailored interactions may influence long-term use; details of these studies can be found in Appendix A.

Moreover, LTE is not only a means of evaluating system performance. It can also be a primary research goal in its own right. Several studies in our review, for example, focused on understanding patterns of user disengagement or identifying the factors that lead to drop-off in daily robot use, particularly in home environments.

Currently, there is no standardized approach for measuring LTE, and the form it takes often depends on the nature of the interaction being studied. In our review, we identified seven common categories of LTE measurement methods, drawn from the diverse practices used throughout the corpus.

- **Self-reported**: This method uses surveys to directly ask users about their levels of engagement over time. Likert scales were often used with this approach, with participants rating questions such as: "How often have you used the robot in the last period?" [96], "I think I would like to use this system frequently" [117], and "I think I could spend a good time with [the robot]" [4]. Some studies instead reported insights from a series of user interviews that asked qualitatively about engagement with a robot. Overall, 28.9% ($N = 13$, [4,58,68,93,96,109,117,118, 121,141,154,156,165]) of studies measuring LTE used self-reported methods.

- **Interaction times**: This method uses temporal measures of human-robot interactions to gauge engagement in the moment and compare these metrics over time. A common method was measuring the duration of robot usage per usage instance, with longer durations as an indication of higher engagement. For instance, Scassellati et al. [3] showed that children played with a robot for a similar average amount of time during the first five sessions of use in comparison to the last five of 23 sessions. Another method was to measure the amount of additional time participants chose to spend with the robot (e.g., [112]). Out of the studies measuring LTE, eight (17.8%) measured it using an interaction time approach [3,91,112,114,119,120,150,157].

- **Annotations**: This method was used on recordings of study sessions to hand annotate user engagement labels during robotic interactions. Trends in the annotations were then compared over time. For instance, Clabaugh et al. [177] annotated a video of children with ASD interacting with a social robot, basing engagement levels on whether a child was "paying full attention to the interaction, immediately responding to the robot's prompts, or seeking further guidance or feedback from others in the room." In total, 13.3% ($N = 6$; [48,108,110,145,148,177]) of studies measured LTE using annotations.

- **Count-based**: This method includes counting the number or rate of certain types of interactions, such as games or activities, that the user performed with the robot over time. For example, Kanda et al. [124] utilized wireless tags to identify individual children who used the robot, and how often, in order to find patterns of drop-off. The *dropout rate* of users between two periods of time can also be calculated from counts of robot versus control use. For example, Barco et al. [43] used this method to show that a robot-supported rehabilitation program had less user dropout than without a robot. A counting technique often used in consumer electronics is reporting the *daily active users* of a device longitudinally. Zhao and McEwen [44] used this method of reporting to find that the daily active users of a Luka robot for reading with children dropped from 20 to six over the course of 180 days. Overall, 11.1% ($N = 5$; [2, 42, 43, 56, 124]) of studies measuring LTE used some count-based approach.

- **Sensor-based modeling**: Several studies employed vision and audio inputs with machine learning methods to estimate user engagement, either in real time or through post-hoc analysis. Commonly extracted features included affect or mood, body posture, vocal tone, and gaze behavior. These predicted engagement metrics were then tracked and analyzed over time. In total, 11.1% ($N = 5$; [48, 90, 122, 159, 166]) of the studies that measured LTE primarily relied on this type of sensor-based approach.

- **Behavioral observation**: One approach to measuring LTE involves live, in-person observation of user behavior during interactions with the robot. This method relies on researchers' subjective interpretations of engagement, often informed by repeated exposure to participants over time. For example, Michaud et al. [144] used direct observation to record how children would proactively assist the robot when the robot did not appear to react correctly to certain stimuli.

- **Mixed**: Seven studies (15.5%) [55, 105, 113, 116, 129, 131, 178] employed a combination of the above approaches to measure LTE.

2.5 Discussion

The volume of long-term HRI research has grown substantially, increasing from 26 published papers between 2003 and 2013 to 94 papers in the more recent decade. This surge reflects a growing commitment to understanding how robots interact with and influence users over extended periods of time.

Our analysis revealed several encouraging trends in long-term HRI research over the past two decades. In particular, we observed a broad representation in age groups, ranging from toddlers to older adults. However, a key gap emerged in the relative scarcity of studies throughout the lifespan—particularly those involving teenagers, who remain underrepresented despite their distinct developmental trajectories and social needs. We explore this research gap and the opportunities it presents in greater detail in Section 2.5.1.[8]

Another positive trend is that most of the studies we analyzed involved the robot operating entirely *autonomously* (75.8%) and *in situ* (85.8%), mirroring the real world environments and contexts in which they will need to function. These trends suggest an increasing alignment between research conditions and the environments in which robots are ultimately expected to operate. Even among the smaller subset of lab-based studies (14.2%), researchers frequently designed the physical and social context to simulate naturalistic environments, such as mock living rooms or classroom setups, helping to elicit user behavior that more closely mirrors real-world interaction patterns.

We also observed a strong correspondence between a study's application domain and its deployment environment. Educational robots were commonly tested in schools, therapeutic robots in rehabilitation or clinical settings, and eldercare robots in residential care facilities. Although this alignment may seem intuitive, it often requires substantial logistical effort and institutional collaboration to place robots in these environments. The consistency in this alignment underscores the field's increasing commitment to ecological validity in long-term HRI research.

In the following sections, we begin by identifying key gaps and emerging opportunities in the field (Sections 2.5.1–2.5.3), informed by the evolution of long-term HRI over the past two decades. We then offer a series of design recommendations (Sections 2.5.4–2.5.6) to guide researchers in designing and evaluating long-term robotic systems in real world contexts. Finally, in the interest of transparency and rigor, we acknowledge several limitations of this review and suggest avenues for addressing them in future research (Section 2.5.7).

2.5.1 Opportunity: Designing for Teenage Participants

Our analysis only identified two long-term HRI papers (1.7% of our corpus) that focused their study on teenagers (ages 13–17). The first study focused on teenagers with ASD and severe developmental disabilities. It analyzed the impact of a robot on their communication skills in secondary school and showed the potential for robots

[8] In parallel with the limited representation of teenagers, our review also revealed a striking absence of long-term HRI studies involving adults with ASD (Section 2.4.2). At the time of writing, only one such study had been published: our own work, detailed in Chapter 6. While this section focuses specifically on teenagers, we underscore that the near-total lack of research on adults with ASD represents an equally urgent gap and a critical direction for future HRI work. Because this issue extends beyond the long-term HRI literature, we examine it more thoroughly in Chapter 3.

to improve communication [5]. The second paper also focused on teenagers with ASD, and the authors found a link between the teenagers' sensory profile and their capabilities to imitate a robot [149]. Given the very limited number of long-term HRI studies involving teenage participants, we identify this as a clear and pressing gap in the literature.

Notably, our review did not uncover any studies that examined how *neurotypical* teenagers engage with robots over extended periods. This presents a wide range of open research questions. For example, how do teenagers interact with robots in home environments, especially in the presence of family members? How might they discuss or share their experiences with peers, and what role does peer perception play in shaping engagement?

In particular, we highlight two domains that warrant deeper exploration: education and mental health. Many long-term HRI studies involving younger children have demonstrated positive educational outcomes, yet little is known about how such benefits may extend to teenagers—an age group for whom identity formation and future planning are especially salient. Robotic systems may offer personalized support or motivation during this formative period. Similarly, mental health represents another vital frontier. The World Health Organization estimates that one in seven adolescents (ages 10–19) experiences a mental health disorder [179]. Given the growing body of research demonstrating the potential of robots to support mental and emotional well-being in other age groups, long-term HRI studies targeting adolescent mental health could yield significant impact and insight.

Another important open question is whether teenagers will adopt and engage with robotic technologies in a manner comparable to other age groups. Research suggests that teenagers have distinct relationships with technology, shaped not only by their developmental stage but also by strong social influences. For instance, adolescents often calibrate their technology use in response to peer norms and perceptions [180], with peer endorsement playing a major role in shaping how and whether technologies are embraced [181,182]. In addition, teens are often the first to adopt new technologies [183], making them a critical population to understand the emerging patterns of use and acceptance.

Given these dynamics, it is especially important to examine how teenagers interact with robots over extended periods of time. Their initial enthusiasm may be driven by novelty, but their sustained engagement is likely to hinge on whether the robot aligns with their evolving identities, social environments, and perceived value. In this way, teenagers can provide a particularly sensitive testbed for understanding the novelty effect—a core concern in long-term HRI as we introduced in Section 2.2.1—given that they are highly attuned to technological trends and are quick to disengage from tools that they find unauthentic, stigmatizing, or socially obsolete. Studying how the novelty effect manifests and fades for this population could yield insights that not only improve robot design for teens but also inform broader principles of engagement across other user groups.

2.5.2 Opportunity: Exploring Workplace Integration

A second area we have identified as needing further exploration is the long-term deployment of robots in workplace settings, particularly focusing on office or business environments (rather than schools or hospitals, as explained in Section 2.3). With this categorization, we identified only four papers in which robots were tested for extended periods in workplaces. The first was an experiment in which a robot acted as an assistant in a collaborative workspace, helping workers with routine day-to-day tasks [88]. The second paper compared different types of robots as they coached employees on mental health, specifically in the workplace [139]. The third paper investigated robots that provide break management at desks [121]. Lastly, the fourth paper explored the social aspects of a fetch-and-carry robot designed to assist motion-impaired users in an office environment [84].

Despite the limited amount of prior research in workplace environments, most people spend a significant portion of their lives at work, dedicating approximately forty hours or more per week to it. Therefore, we believe that it is important to study how robots can assist and interact with us in the workplace. Many open questions remain about how robots might impact work environments—will they be accepted and incorporated organically by employers and employees? Will robots increase or decrease productivity? How will they affect the well-being of employees?

In addition to office-like settings, industrial and factory environments represent another prominent category of workplaces. Although numerous studies have examined collaborative and manufacturing robots in short-term or single-session contexts, we did not identify any comprehensive long-term HRI investigations within this subdomain. This absence is notable given the scale and technological relevance of the sector. For instance, manufacturing accounts for approximately 8% of the workforce in the United Kingdom alone [184], and robots are already widely deployed in these settings, with their presence continuing to grow rapidly [185].

The lack of long-term studies in this domain leaves important human-robot interaction questions unresolved. Chief among these are whether robots will be accepted by workers over extended periods—particularly amid rising concerns about job displacement and automation [186]—and how these systems can cultivate trust, coordination, and effective working relationships with human collaborators. Understanding long-term dynamics in manufacturing contexts is essential to ensure not only technical integration but also social acceptance and sustainable deployment.

Finally, future research should broaden its scope to include the diverse range of workplace environments that remain underexplored in long-term HRI studies—such as restaurant kitchens, service industry settings, construction sites, and other non-office, non-industrial domains. Each of these contexts presents unique social, spatial, and operational dynamics that can shape how robots are perceived, integrated, and used over time. A key question for future work is how robots can meaningfully contribute to these settings in the long term—not only through functional assistance, but also by enhancing worker well-being, safety, and collaboration.

Moreover, while students and patients have been frequently studied as primary participants, gaining insight into the experiences and perceptions of other key workplace stakeholders—such as teachers, aides, and administrators in schools, or doctors, nurses, and support staff in hospitals and eldercare facilities—can provide a more complete and context-sensitive understanding of robots' roles in human systems. This expanded perspective is particularly important given the growing body of evidence suggesting that robots can contribute to improved workplace mental health [139], support healthy work practices [121], and even enhance productivity [88].

2.5.3 Opportunity: Standardizing Long-Term Study Metrics

Throughout our analysis of the papers included in this review, we observed frequent use of widely adopted HRI survey instruments such as the Godspeed Questionnaire Series [187] to assess participant perceptions of robots, the Robotic Social Attributes Scale (RoSAS) [188], which evaluates judgments of a robot's social characteristics, and the Negative Attitudes toward Robots Scale (NARS) [189], which measures aversive predispositions toward robots. These standardized tools have provided valuable common ground for comparing results across studies with different robots, participant groups, and interaction contexts.

However, many of these surveys were not originally designed or validated for use as repeated measures over time. In long-term HRI studies, these instruments are often administered multiple times to assess evolving perceptions, but their psychometric stability under such longitudinal conditions remains uncertain. For example, the Godspeed questionnaires were initially developed to guide design decisions during robot prototyping, not to track attitudinal change across extended interactions. Indeed, the authors themselves caution that human perception of robots is "not stable" [187] and is likely to shift as users become more familiar with a robot.

Given the increasing prevalence of longitudinal studies in HRI, it is worth asking if the field now requires updated or entirely new instruments that are explicitly designed and validated to measure changes in perception, trust, acceptance, and engagement over time. Such tools could offer greater reliability and interpretability in long-term settings, ensuring that researchers capture meaningful trends rather than measurement artifacts.

Thus, our final suggested opportunity for future research lies in the development and validation of standardized measurement tools tailored specifically for long-term HRI. We argue that existing instruments are limited in their ability to capture the dynamic and evolving nature of human-robot relationships—particularly in relation to the persistence of the novelty effect and the challenge of sustaining user engagement over extended periods. While tools such as the Godspeed, RoSAS, and NARS questionnaires have been invaluable in establishing foundational insights, our review did not identify any survey instrument that systematically addresses these long-term dynamics in a standardized manner.

To advance the field, we propose two complementary directions. First, researchers may consider extending and updating existing tools to explicitly incorporate constructs relevant to longitudinal interactions, such as relationship progression, habituation, and sustained trust or interest. Second, future work should empirically validate the use of these commonly adopted instruments in repeated-measures contexts to ensure their reliability and interpretability over time.

Another important direction involves developing standardized approaches to assessing specific long-term interaction qualities—particularly the novelty effect and LTE. For example, there is currently no consensus on how to determine when the novelty of a robot has "worn off," nor is there a standardized method for measuring sustained engagement over time. As noted in our review (Section 2.4.4), the studies that measured LTE employed a wide range of methods, including behavioral observations, surveys, and sensor-based techniques. While variation is expected—given differences in research goals, participant populations, environments, and application domains—establishing more consistent measurement frameworks within similar study types would offer substantial benefits. As long-term HRI continues to grow as a field,

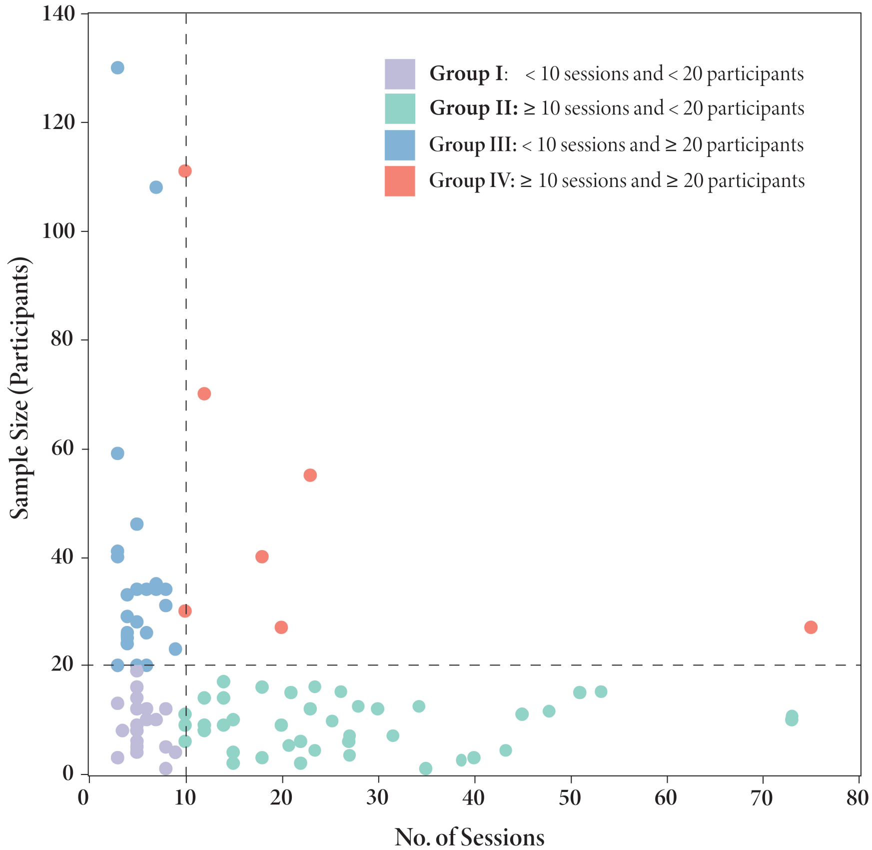


Figure 2.10: Distribution of Session-Based Studies by Sessions & Sample Size. Four primary categories emerge: studies with < 10 sessions and < 20 participants (*Group I*, lower left, purple), studies with ≥ 10 sessions and < 20 participants (*Group II*, lower right, green), studies with < 10 sessions and ≥ 20 participants (*Group III*, upper left, blue), and studies with ≥ 10 sessions and ≥ 20 participants (*Group IV*, upper right red). Four studies [5–8] were considered outliers and are excluded from this plot for clarity.

having standardized benchmarks and methods will not only facilitate cross-study comparisons but will also accelerate the development of more robust and impactful long-term robotic interactions.

2.5.4 Recommendation: Determining Core Study Features

For researchers planning a long-term HRI study, determining the appropriate duration of the study is a critical decision with implications for scheduling, funding, participant recruitment, and technical feasibility. Our analysis suggests that this decision is often closely linked to both the number of participants involved and the nature of the data being collected—whether quantitative, qualitative, or a combination of both.

Figure 2.10 shows four primary categories, as determined by observation. *Group*

I encompasses session-based studies with fewer than 10 sessions and fewer than 20 participants ($N = 35$ studies; 40.7%). These studies may be motivated by a blend of practical considerations. Here, resource limitations or the inherent complexity of sustained HRI interactions could steer researchers toward concise study durations and small participant pools. *Group II* features studies with 10 or more sessions but fewer than 20 participants ($N = 14$ studies; 16.3%). With a relatively small number of participants but many sessions, such studies emphasize longitudinal depth or within-subject analysis, not statistical power or broad participant representation. In the third category (*Group III*), characterized by studies with fewer than 10 sessions and 20 or more participants ($N = 29$ studies; 33.7%), researchers can seek to gather insights from diverse participants despite the comparative brevity of the study. Lastly, the smallest cluster (*Group IV*) comprises studies with 10 or more sessions and 20 or more participants ($N = 8$ studies; [6–8,44,85,140,143,163]). A high number of participants for a long amount of time is ideal from a research perspective, but it is a practical and logistical challenge. In this group, researchers may be driven by the desire for a comprehensive exploration within a more specialized context.

We present these categories not as a rigid classification system, but as a reflective tool to help researchers consider where their study might fall—and where they aspire for it to fall. When faced with the necessary question, *"How long should my study be?,"* researchers should consider study length in relation to participant type and measurement goals. For instance, a qualitative study with a small number of users may benefit from longer durations to yield meaningful insights; however, a quantitative study aiming for statistical significance may require a larger sample size, thereby favoring shorter interactions for feasibility. We encourage researchers to consult our corpus as a practical reference for how prior studies have navigated these trade-offs across different domains, settings, and participant populations.

2.5.5 Recommendation: Sustaining Engagement With Novel Behaviors & Personalization

As highlighted in Section 2.4.3, robots that demonstrate a range of varied and responsive behaviors tend to sustain user engagement more effectively than those with rigid or repetitive interaction patterns. One compelling strategy to achieve this variability is through personalization: that is, tailoring the robot's behavior to align with the preferences, needs, or learning styles of individual users [48,113]. By building a model of the user's behaviors and adapting accordingly, the robot can deliver interactions that feel more relevant, responsive, and human-like. Prior research has shown that personalized robotic systems can lead to enhanced learning outcomes [162], stronger engagement, and increased rapport between the user and the robot [190], all of which are critical for successful long-term interaction.

Our analysis shows that only a minority of the studies in the corpus (39.2%) incorporated robot adaptation or personalization during interactions, although this number is steadily increasing over time. We strongly encourage future researchers to integrate adaptation and personalization mechanisms into their robotic systems where appropriate, as these features are often critical to sustaining long-term user engagement, acceptance, and continued use. There are a wide range of strategies to introduce adaptation in long-term HRI. For example, some studies have maintained a continuous backstory for the robot across sessions to create a sense of narrative continuity [3]. Others have designed robots capable of skill progression, allowing them to display increasingly complex behaviors over time [130], or have incorporated multiple activities to vary interaction and prevent monotony [190]. Personalization can also be achieved in diverse ways, such as remembering and reusing user-specific information like names [141], modeling individual skill levels to tailor tutoring behaviors [49], recognizing the user's current context or activity [116], or learning and responding to user preferences over time [104]. These techniques illustrate the breadth of opportunities for creating socially responsive, user-aware robots capable of fostering deeper and more meaningful long-term interactions.

We encourage researchers to consult this review as a resource for identifying prior long-term HRI studies that align with their intended domain, population, and interaction context—both as methodological inspiration and as a basis for comparison. For those seeking more focused insights into adaptive and personalized human-robot interactions, several dedicated reviews explore the technical and design challenges in this space, including works by Gasteiger et al. [191], Ahmad et al. [176], and Hellou et al. [175]. At the same time, this area remains ripe for innovation. With ongoing advances in artificial intelligence, machine learning, and sensing technologies, there is a growing opportunity to define novel methods of adaptation and personalization tailored to long-term use.

Crucially, the success of such methods is highly dependent on the specific context of interaction. For example, personalizing a robot that engages with multiple users in a shared environment (such as a school) introduces identity management challenges, especially when sessions are brief or users frequently change. Group-based interactions introduce another layer of complexity, as certain personalized behaviors may be socially appropriate within one subgroup but awkward or exclusionary in a mixed setting. Domain-specific considerations also matter: in *Physical Health* contexts, participants may have differing physical capacities that influence how they can interact with a robot, thereby affecting the kinds of adaptations that are both possible and meaningful. Similarly, a robot with a limited interaction channel (such as the nonverbal Paro) offers fewer pathways for personalization compared to a multimodal platform like NAO, which supports speech, gesture, and visual feedback.

These examples illustrate that there is no universal approach to personalization and adaptation in long-term HRI. Instead, the design must be carefully shaped by the robot's capabilities, the participant profile, the social and physical environment, and the goals of the interaction. We therefore urge researchers to clearly articulate in their work which aspects of their personalization and adaptation strategies are context-specific and which might be generalizable to other domains or populations. Doing so will enrich the field's collective understanding of how adaptive robotic systems can scale, translate, and evolve across long-term, real-world deployments.

2.5.6 Recommendation: Reporting the Full Data & Context

Many of the papers analyzed in this review lacked key information or statistics necessary to characterize the long-term nature of robotic interactions. In session-based studies, the most frequently omitted details were the number of sessions, the duration of each session, and the total or average time the participants spent interacting with the robot. These metrics are critical for the HRI community in assessing how long-term exposure influences outcomes such as user engagement, habituation, and dissipation of novelty effects. Where possible, we estimated missing values, such as average minutes of interaction, for the purposes of this review. However, we strongly encourage future work to report these metrics consistently and transparently, as they are essential to allow meaningful comparisons between studies and to advance a cumulative understanding of long-term HRI.

Several of the studies in our corpus were categorized as *free-use* deployments, in which users had the freedom to decide when and how to interact with the robot. These studies offer valuable insights into how people engage with robots in naturalistic settings such as homes and schools, free from the constraints of tightly controlled experimental protocols. They are particularly useful for understanding which user demographics are most likely to engage with the robot and under what contextual conditions these interactions occur. Moreover, free-use studies offer a unique opportunity to observe how engagement patterns evolve or diminish over time. While many of these studies reported the total duration the robot remained in users' homes, we found that several lacked crucial contextual details—such as the frequency and duration of daily interactions, the identity of users or family members involved, and the specific times of day or scenarios in which the robot was used most often. We encourage future work in this area to systematically capture and report these behavioral patterns, as doing so can greatly enhance our understanding of real-world robot usage and inform the design of more engaging long-term systems.

One aspect that was rarely reported across studies was whether participants received compensation for their involvement. This omission is particularly important in the context of long-term studies aiming to evaluate sustained user engagement or compliance. Participant motivation can significantly impact study outcomes: some individuals may continue to interact with the robot due to genuine interest and engagement, while others may be driven primarily by incentives or a sense of obligation to the research team. Without transparency in compensation, it becomes difficult to interpret whether long-term engagement reflects authentic interest or external motivators. We recommend that all long-term HRI studies explicitly state their incentivization strategies, including whether and how participants were compensated, so readers can better evaluate potential confounding factors and the validity of user engagement outcomes.

2.5.7 Review Limitations

There are several limitations to our review. Despite employing a rigorous search methodology, it is likely that some relevant studies were unintentionally excluded. In particular, our focus was primarily on human-robot interaction conferences and journals; we did not conduct an extensive search of more traditional robotics venues, which may contain additional long-term interaction studies. A second limitation stems from our exclusion criteria: we omitted studies in which the robot did not engage with the same user across time (e.g., museum deployments). While our aim was to focus on sustained, longitudinal user-robot relationships, there is also considerable value in examining how robots interact with diverse, changing user populations over extended deployments. Third, our analysis emphasizes the primary characteristics of each study, which may obscure important nuances—such as studies that span multiple populations, domains, or include multiple phases that reflect different temporal dynamics. Finally, we had to estimate certain values for a number of papers, including interaction duration and average session length, due to incomplete reporting. These estimates, while necessary for comparative analysis, may not fully reflect the original study design or outcomes.

2.6 Summary

The synthesis of 120 long-term HRI studies presented in this review highlights both the rapid expansion and the increasing complexity of long-term social robotics research. By adopting a broad perspective, we traced how the field has evolved over the past two decades to identify key trends in robot autonomy, real-world deployment, participant demographics, and evaluation methodologies.

These insights directly inform the broader goals of this dissertation. First, the growing focus on in-situ, long-term deployments underscores the urgent need to design robots capable of sustaining meaningful social engagement over extended periods—particularly in dynamic, real-world environments such as homes, schools, and care facilities. Second, recurring interaction patterns across studies offer promising models for how robots can scaffold learning, support social and emotional development, and adapt to individual users through mechanisms such as personalization. Third, persistent gaps—such as uneven age representation across the lifespan, the scarcity of workplace-oriented systems, and the absence of standardized tools for assessing long-term engagement and success—highlight the need for more inclusive, contextually grounded, and methodologically robust approaches to HRI research.

Together, these findings shape the core motivations of this dissertation. By addressing critical gaps and building on emerging best practices in long-term HRI, this work advances the design, development, and deployment of socially intelligent robots that can meaningfully support users across a range of life stages, contexts, and social goals. Our later chapters (Chapters 4–6, 8, 9) directly extend the literature. Across five studies, we examine how robots can be tailored to specialized populations, integrated into real-world environments, and evaluated through sustained long-term interaction.

Chapter 3

Robots for Autism Therapy

The previous chapter examined how the field has approached extended interactions between humans and robots. We highlighted emerging trends, foundational design assumptions, strategies for sustaining engagement with robots, and persistent research gaps. These insights directly inform the aims of this dissertation across its three central dimensions: the *design* of robots for social interaction, their technical *development*, and the contextual factors that enable their successful *deployment*. While the prior chapter surveyed a broad range of application domains, from entertainment to physical health, this chapter focuses specifically on one of those domains: robot-assisted autism therapy. Here we review more than 300 studies involving the use of socially assistive robots in autism interventions—not only because autism has been a prominent focus within robotics research, but also because it offers a uniquely rich testbed for examining the mechanisms underlying socially mediated learning. Core diagnostic features of autism—including difficulties in social communication, emotional regulation, and adaptive behavior—closely align with the domains where robots are believed to offer the most therapeutic value. As such, the autism literature provides critical insights into both the potential and limitations of robot-based interventions.

3.1 Introduction

Formally known as Autism Spectrum Disorder (ASD), autism encompasses a broad range of neurodevelopmental conditions, marked by significant variability in communication styles, cognitive profiles, sensory processing, and daily functioning. The operational criteria have evolved over time, sometimes ahead of fully conclusive scientific consensus and reflecting shifting clinical perspectives [192, 193]. Yet, core diagnostic hallmarks have remained consistent: persistent difficulties in social communication and interaction, alongside restricted, repetitive patterns of behavior, interests, or activities [193, 194].

Currently, there is no cure for autism,[1] but a range of behavioral treatments have been shown to meaningfully improve quality of life and support greater independence. Early intervention programs, in particular, aim to target foundational social and adaptive skills during critical psychodevelopmental windows [195, 196] in order to maximize the potential for lasting, long-term impact [197]. However, these programs demand sustained time, expertise, and involvement from families, clinicians, and educators—making equitable access a persistent challenge [198]. These challenges are further intensified by the profound heterogeneity of the autism spectrum, which necessitates highly individualized care. However, such personalized models are difficult to implement at scale within institutional systems that often default to standardized protocols (e.g., in public schools [199], child welfare models [200], or healthcare [201]).

To supplement the level of human involvement required for personalized and readily available care, some approaches have explored the use of non-human partners to facilitate human-human social interaction, such as in pet-assisted therapy [202, 203]. Digital tools such as computer-assisted programs and virtual reality platforms have also shown potential to support engagement and skill development in individuals with ASD [204, 205]. However, there remains limited research on the specific mechanisms that make these facilitative interactions effective and on the conditions necessary to generalize the benefits to real-world engagement with human partners.

Robots, particularly socially assistive robots (SARs), extend this line of inquiry by offering physically embodied, interactive systems that can engage users in structured, socially meaningful ways. Unlike virtual agents or passive media, robots occupy physical space, respond dynamically to user behavior, and can model or reinforce key social behaviors through real-time interaction. As a result, SARs hold unique promise as therapeutic tools that not only simulate aspects of human engagement, but also actively support the acquisition and generalization of social skills across diverse settings. Research on SARs for autism shows increased engagement, improved attention regulation, and more appropriate social behavior such as joint attention and spontaneous imitation when robots are part of the interaction [20, 21].

[1] We acknowledge this phrase is common in clinical discourse but controversial within the autism community. While it underscores autism's permanence as a neurodevelopmental condition, it is also critiqued for pathologizing autism identity and conflicting with neurodiversity perspectives that emphasize acceptance and accommodation.

Scope of This Review

This review aims to describe the current state-of-the-art in robots for ASD therapy and, in doing so, to make the results accessible to a broad interdisciplinary audience. The field contains many studies with different methods and goals, but the projects can generally be divided into three connected but discrete phases: designing the intervention goals and structure; engineering the robot's physical form and behavior to deliver those goals; and evaluating the outcomes of the robot-assisted intervention. In particular, **intervention design** (Section 3.3) focuses on identifying the social, cognitive, or behavioral goals the robot is meant to support: What skills should the robot support (e.g., joint attention, emotional regulation, social reciprocity), and through what types of activities or interaction sequences? Should the intervention target individual users or support peer interaction? How should goals be adapted for different age groups or cognitive profiles? **Robot development** (Section 3.4) addresses the questions of form and function: What appearance, movement, or expressive modalities will best support the intervention? Should the robot display affect through facial features, body motion, or vocal tone? Will it need arms to gesture, a head to orient, or mobility to reposition within the environment? How autonomous should it be and how will it sense, interpret, and respond to user behavior in real time? Finally, **evaluation** (Section 3.5) considers whether and how the system achieves its intended outcomes: Are target behaviors improved over time? Does the robot support engagement, generalization, or retention of skills? How do users, families, and clinicians perceive its usefulness and appropriateness? These phases are often iterative and overlapped, but together they form a common structure for designing, developing, and deploying robots for autism therapy.

This review draws on an extensive collection of peer-reviewed studies that involve interactions between at least one robot and at least one individual with ASD. We include studies that present a robot that is physically present and play an active role in social interaction; studies featuring virtual agents, screen-based representations, or robots limited to purely mechanical or non-social assistive functions are excluded. Studies are included only if they explicitly state that participants have a formal diagnosis of ASD, verified through clinical evaluation or standardized diagnostic instruments such as the DSM-5 or ADOS. Given our inclusion criteria, the final corpus consists of 304 papers and is listed in Appendix B.

Numerous reviews have examined the use of robots in autism therapy, reflecting growing interdisciplinary interest on the topic across robotics, psychology, and clinical science. Among these, two seminal reviews published in 2012 stand out for their foundational influence. Scassellati et al. [20] offered a robotics-centered perspective that emphasizes system design, behavior modeling, and early technical challenges. Diehl et al. [206] assessed clinical utility, critically examining the therapeutic validity and evidence base of robotic interventions to outline methodological gaps in the literature. Both reviews concluded that robots showed promise for eliciting social behaviors and engagement in children with ASD, but that most studies were preliminary, with small samples, minimal experimental rigor, and very limited demonstration of clinically meaningful outcomes. As a result, these reviews formalized the research agenda for the following decade and identified critical gaps in standardization, evaluation, and clinical integration.

Since 2012, the field has evolved considerably. This review builds on the foundation laid by these earlier reviews but differs in scope and focus: it systematically synthesizes research from its origin in 2000 across two decades to 2024, emphasizes embodied robot–human interaction with individuals formally diagnosed with ASD, and foregrounds the therapeutic, technical, and methodological evolution of the field over time. It also seeks to bridge robotics and clinical perspectives, offering an integrated view of progress to date while identifying persistent gaps and future directions.

3.2 Field Growth and Trends

The studies included in this review span from 2001 to 2024, covering more than two decades of research on robot-assisted interventions for individuals with autism. The distribution reveals substantial growth in publication volume over time, with the most pronounced expansion occurring in the last decade. Since the prior literature reviews conducted in 2012—which together captured 21 studies, a subset of the 55 studies captured in this present review—an additional 249 studies have been published within this recent decade (from 2013 to 2024).

In the early 2000s (2001–2004), only four studies were published [207–210], reflecting an exploratory phase of socially assistive robotics and early proof-of-concept interventions. This number grew modestly in 2005–2008, with 15 studies, as researchers began introducing structured pilot evaluations and early humanoid platforms.

The field gained momentum between 2009 and 2012, during which 36 studies were published, coinciding with wider access to off-the-shelf, programmable robots and interdisciplinary interest. A more substantial expansion occurred in 2013–2016, with 64 studies—almost doubling the output of the previous four-year period. This

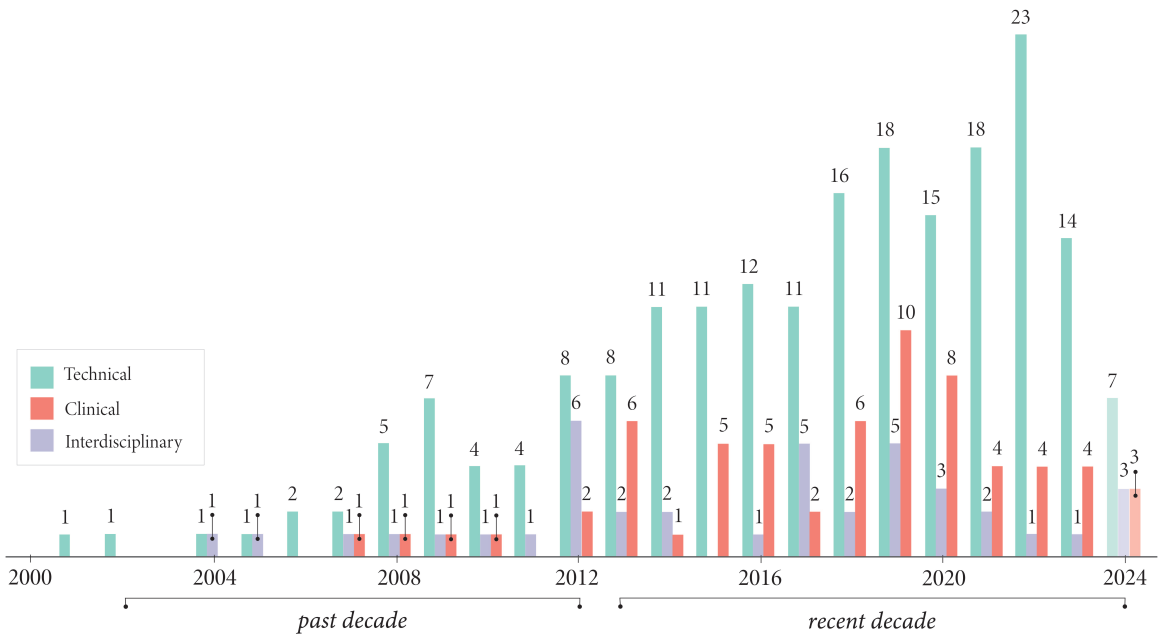


Figure 3.1: Annual Publication Count by Venue Domain. This figure shows the number of published studies per year categorized by broad venue domain: technical (blue), clinical (red), and other or interdisciplinary (purple). The overall trend reflects substantial growth over the past two decades, with the most significant expansion occurring in the last decade. We note that the paper corpus was collected in December 2024. As a result, publication counts for 2024 may not fully reflect that year's conference proceedings or late-year journal publications.

surge reflects both technological advances (e.g., improved sensors, greater autonomy) and an increased focus on social, emotional, and language-based interventions.

The most significant acceleration occurred in 2017-2020, which saw 101 studies and marked a shift toward richer behavioral targets, more autonomous systems, increased cross-sector collaboration, and preliminary real-world deployments.

In contrast, 2021–2024 saw a decline to 84 studies, representing a 17% decrease relative to the previous four-year period. This drop may reflect research constraints imposed by the global COVID-19 pandemic, a shift in publication priorities, or a maturation of the field wherein feasibility studies are giving way to fewer but more rigorous time-intensive deployments. Nonetheless, the recent output remains historically high and underscores the field's sustained relevance and evolving focus on robot-based therapies for ASD.

3.2.1 Contextual Shifts Affecting the Research Landscape

As summarized in the previous section, the development of robots for ASD therapy has evolved in parallel with several key contextual shifts that have shaped the field's growth and research output. We organize these shifts into four major themes: changes in the diagnostic framework, increased prevalence and funding opportunities, the impact of the global COVID-19 pandemic, and advances in technological infrastructure.

New Diagnostic Framing

In 2013, the publication of the DSM-5 redefined autism diagnosis by consolidating previously distinct subtypes—Asperger's Disorder, Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS), Childhood Disintegrative Disorder (CDD), and, previously, Rett Syndrome—under a single ASD umbrella category.[2] This reclassification emphasized a dimensional, spectrum-based understanding of autism and introduced severity levels based on the amount of support required (Levels 1–3). It also reorganized diagnostic criteria into two core domains: deficits in social communication and restricted, repetitive patterns of behavior. For researchers, this shift had significant implications: it required revisions to participant inclusion criteria, encouraged recruitment of more heterogeneous samples, and complicated cross-study comparisons where earlier diagnostic labels had been used.

Importantly, these diagnostic changes did not create new participant subgroups but rather drew attention to populations that had long been underrepresented in research. Before DSM-5, the possibility of formally diagnosing multiple co-occurring conditions was limited, which obscured the high rates of overlap with other neurodevelopmental and mental health conditions [211, 212]. Similarly, females with autism were often overlooked due to behavioral and clinical presentations that diverged from established male-centric diagnostic profiles [213, 214]. Minimally verbal individuals, too, were historically excluded from research participation, in part because of limited tools and infrastructure for accommodating their needs [215]. In this sense, it is our recognition and methodological responsiveness—not autism itself—that has changed.

These diagnostic changes occurred alongside a broader transformation in the clinical and societal understanding of autism. The neurodiversity movement, which gained

[2]The Diagnostic and Statistical Manual of Mental Disorders (DSM) is the primary classification system used by clinicians and researchers in the United States to diagnose mental and developmental disorders. DSM-5 refers to its fifth edition.

momentum in the 2010s and 2020s, reframed autism from a deficit-based model to a difference-oriented perspective. This perspective has influenced robotics research, as reflected in recent discourse among roboticists [216], prompting a shift away from deficit-based approaches such as proposing robot-led therapy for "correcting" behavior. This shift advocated for more participatory design approaches and stakeholder-informed methods that actively involve people with ASD in the design and evaluation of SARs.

Rising Prevalence and Funding Shifts

Estimates of autism prevalence have shifted significantly across these two decades. While ASD was once considered relatively rare, with data in the early 2000s suggesting a prevalence of 1 in 150 children [217], more recent figures estimate that 1 in 36 children meet diagnostic criteria [218]. Much of this rise can be attributed to changes in diagnostic practices, including increased screening, earlier detection, and broader definitions introduced in the DSM-5. Simultaneously the rising prevalence drew heightened attention from policymakers, funding agencies, and the research community.

This attention culminated in a wave of large-scale long-term funding initiatives for autism research, particularly in the early 2010s. In 2012, the U.S. National Science Foundation (NSF) awarded two highly visible Expeditions in Computing grants (each $15 million in size) focused explicitly on technologies for autism. At the same time, similarly ambitious programs emerged across Europe, reflecting a shared global priority.[3] These grants not only enabled research teams to pursue more ambitious and longitudinal goals but also provided stability and interdisciplinary capacity beyond the traditional 2–3 year grant cycle. This inflection point had a profound effect on the field: prior to 2012, our corpus included only 55 studies. In the decade that followed, that number grew nearly fivefold to 249 studies. The influx of funding may have spurred increased research productivity, fostered interdisciplinary collaboration (as discussed in Section 3.2.2), and enabled more ambitious and methodologically diverse experimentation.

[3]For example, in the United States, two NSF Expeditions in Computing grants were awarded: one led by Yale University focused on socially assistive robotics (Award 1139078; [219]), and another led by Georgia Tech centered on computational behavioral science (Award 1029679; [220]). In Europe, major multi-institution projects such as ROBOSKIN (FP7 ICT Grant 231500; [221]), ALIZ-E (FP7 ICT Grant 248116; [222]), and DREAM (H2020 Grant 645753; [223]) encouraged dedicated research on robots for autism therapy.

Global Pandemic Disruption

The COVID-19 pandemic (2020–2022) significantly disrupted both clinical services and robotics research. In-person therapy sessions were suspended or transitioned to virtual formats, interrupting access to behavioral interventions that are especially time-sensitive for young children. The pandemic exposed the fragility of existing therapeutic infrastructure, particularly its reliance on in-person delivery models, and prompted renewed interest in technologically mediated solutions to supplement human-provided care. Robots presented opportunities to deliver or sustain therapy when human providers are unavailable, overburdened, or difficult to access [224].

However, SAR research relies heavily on in person, embodied interaction, and was therefore affected by the pandemic. Human-robot interaction studies were paused, planned deployments were delayed or canceled, and longitudinal data collection was interrupted. These limitations forced the field to reimagine its methodologies and accelerated interest in telepresence robots, remote sensing, and hybrid delivery models that could function under constraints of physical distancing. The traditional paradigm of robotics research, which involved inviting participants into a controlled laboratory or clinical setting for supervised interaction with a robot, was challenged. Researchers were therefore compelled to explore how SAR systems might function autonomously and adaptively in the dynamic, unstructured environment of participants' homes with minimal or no experimenter supervision. During the pandemic, even the basic requirement of running a study necessitated grappling with longstanding technical challenges (including reliability, adaptability, and remote operability) that had previously been considered aspirational or secondary [225].

In light of pandemic constraints, several studies reimagined SAR deployment models; three such examples are outlined here. Katsanis et al. [226] developed a compact, low-cost robot using off-the-shelf hardware and 3D-printed components, prioritizing affordability, replicability, and ease of home deployment under access restrictions. To facilitate human-directed therapy, Fischer et al. [227] (conducted before the pandemic but immediately impactful during) demonstrated the utility of a telepresence robot equipped with video, audio, and mobility capabilities for remote behavioral consultation in ABA settings. A remote therapist could navigate the robot within the environment to observe therapy sessions, assess prompt dependency, and evaluate patient outcomes without being physically present. At the level of full-system deployment, Ramnauth et al. [29] described the development of an autonomous, low-touch SAR platform designed for fully contactless delivery: the system could be dropped off at a participant's home, set up entirely by the user, and maintained remotely by researchers with minimal intervention.

Technological Advancements and Infrastructure Growth

Early research (such as that captured in the two 2012 reviews) on social robots for autism therapy was conducted almost exclusively by a small number of research groups. These early studies required close collaboration with clinicians and educators, as well as ethical oversight and specialized recruitment methods to work with a protected and highly heterogeneous population. At the same time, building a robot capable of engaging in social interaction required significant custom hardware and low-level programming expertise. Over the past decade, however, major advances in both robotics and AI have transformed the landscape. The emergence of commercially available, socially expressive platforms and open-source toolkits has enabled broader participation, faster prototyping, and greater reproducibility.

The mid-2010s marked a turning point in the field with the proliferation of off-the-shelf, socially expressive robotic platforms such as NAO [228], Pepper [173], and QTrobot [229]. Unlike earlier systems that required extensive custom engineering, these platforms offered modular software environments, standardized APIs, and user-friendly development toolkits. Their articulated bodies, expressive faces, and programmable behaviors enabled a wider range of social cues, such as gaze shifts, gestures, and speech, to be integrated into therapy sessions. As a result, these robots became widely adopted in autism research and intervention studies, particularly to model social behavior, deliver prompts, and facilitate structured interaction. Their relative affordability, ease of use, and growing developer communities significantly lowered the barrier to entry, allowing new first-time researchers without robotics expertise to develop and deploy SAR-based interventions more rapidly and at greater scale.

Over the past decade, researchers have increasingly advocated for open-source development and the public release of large-scale, multimodal datasets. Open-source software packages such as openSMILE (for audio feature extraction) [230] and OpenFace (for facial expression and gaze tracking) [231] have become standard tools in the field, allowing researchers to analyze vocal prosody, eye gaze, facial action units, and head pose in real time. Robots could now recognize and adapt to these fine-grained social signals from users, allowing more responsive and context-aware interactions. These capabilities were especially important in autism interventions, where sensitivity to subtle changes in user behavior could support better prompt timing, adaptive feedback, and more naturalistic participation. With this, there has been growing attention to the creation of datasets that reflect the unique behavioral patterns of children with ASD, as most available datasets and tools reflect neurotypical adult behavior. For instance, the Engagnition dataset [232] includes multimodal behavioral and physiological recordings from children with autism during robot-mediated interactions.

Whether driven by updated diagnostic frameworks, the rising prevalence of autism worldwide, the increased access to off-the-shelf technology and platforms, or shifts in research protocols due to the pandemic, SAR-based autism interventions have become increasingly global in scope. While early studies were almost exclusively concentrated in North America and Western Europe, our corpus reveals a notable expansion in geographic diversity. Countries such as Japan (30 studies), Malaysia (11), Iran (9), Hong Kong (9), Kazakhstan (8), India (7), and Brazil (5) now contribute substantially to the research landscape. This shift reflects growing international interest in robot-assisted therapy and suggests that SAR interventions are increasingly being developed and studied within a wider range of cultural, economic, and clinical contexts.

3.2.2 Publication Venues and Disciplinary Domains

Despite a growing number of interdisciplinary collaborations, SAR research for autism therapy remains predominantly at the intersection of two disciplinary domains: robotics (typically aligned with computer science and engineering) and clinical psychology or related health sciences. Each domain maintains its own research paradigms, methodological priorities, and publication conventions.

In robotics, high-impact findings are often disseminated through competitive, peer-reviewed, annual conference proceedings, such as the *IEEE International Conference on Robotics and Automation (ICRA)* and the *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. These venues typically feature short-format papers (6–10 pages) and emphasize technical innovation such as in robotic hardware, software architecture, or data and perception pipelines. User testing in robotics frequently center on fine-grained qualitative and quantitative data collected from a small number of participants. From our corpus, it is not uncommon for studies to include detailed, second-by-second analyses of system performance or user behavior based on data from a single participant or fewer than ten. The emphasis is more often placed on demonstrating the robot's feasibility within the target user setting, rather than on establishing therapeutic efficacy or the generalizability of outcomes.

In contrast, clinical psychology studies typically appear in longer manuscripts (10–30+ pages) in monthly or quarterly journals that are peer reviewed and also highly competitive. These studies tend to be large-scale experiments that involve hypothesis-driven controlled trials, are statistically powered to detect group-level effects, and evaluate therapeutic efficacy through validated outcome measures. Journals such as the *Journal of Autism and Developmental Disorders (JADD)*, *Autism Research*, *Molecular Autism*, *Research in Autism Spectrum Disorders*, and *Autism* represent primary outlets in this domain.

These differences between what clinicians value and what roboticists value are important. To illustrate, a robotics paper may emphasize that a robot's emotion classifier achieves 85% accuracy or that its engagement model outperforms a baseline in predicting attention. While these are meaningful technical milestones, they may have limited clinical relevance if not linked to real-world improvements in a child's behavior or functioning. From a clinical perspective, the more pressing question might be: Does detecting a child's emotional cues with 85% accuracy translate into improved social participation? If a child is still not initiating interactions with peers, then even the most accurate classifier may not meaningfully impact therapeutic outcomes. The most impactful studies would connect the two—e.g., showing that an emotion classifier enabled the robot to keep the child engaged longer, which correlated with the child learning more, which lead to improvements in the child's experiences outside of the specific human-robot context.

The two disciplines also have different timelines and incentive structures. In robotics, the field rewards rapid iteration and technical novelty, often leading to the publication of new systems within months of their development—even before the system has undergone extensive user testing. In contrast, clinical research typically unfolds over much longer timeframes, requiring extended recruitment, standardized assessment, and ethical oversight, spanning several years. This temporal misalignment is exacerbated by the short lifecycle of robotic platforms: many commercial or research-grade robots have become outdated or unsupported within 3–5 years of their release,[4] making it difficult to sustain research beyond initial pilot studies. As

[4]For example, Pepper ceased production in June 2021 and lost active software updates following Aldebaran's insolvency in early 2025, while NAO's last major hardware revision was in 2018 and its developer entered receivership in mid-2025. Other historic examples include MIT's Jibo, which became unsupported within two years of its 2017 launch. Similarly, support, cloud connectivity, and SDK access for the Cozmo and Vector robots were discontinued in 2019 following the closure of their parent company, Anki.

a result, clinically validated interventions may be tied to platforms that no longer reflect the state-of-the-art, while state-of-the-art platforms may lack the empirical validation required to demonstrate therapeutic use.

At a general level, the majority of the studies (200 studies; 66%) in our corpus were published in technical venues, where 64 studies (21%) appeared in clinical venues. A smaller subset of 40 studies (13%) were in more balanced or mixed-interdisciplinary venues. Almost every year,[5] the number of technical studies has consistently outnumbered that of clinical or other interdisciplinary studies by a factor of two or more.

Following 2012, at which point only six clinical studies had been published [233–238]), the field experienced notable growth in both the research output and the disciplinary breadth. Of the 249 studies published from 2013 onward, 23% (58 studies) appeared in clinical venues and 11% (27 studies) in interdisciplinary outlets. This marked the first sustained presence of clinical studies featuring a robot for ASD therapy, indicating a shift toward more clinically engaged audiences and growing cross-disciplinary alignment.

While previous periods showed a steady increase in overall research output, the 2021–2024 period saw a surprising plateau in publication volume. Technical papers, however, continued their gradual ascent, with the most substantial growth in 2013-2016 (from 23 studies in the previous 4-year period to 42), to comparatively less sustained growth in 2021-2024 (from 60 studies in the prior period to 62). Clinical publications followed a different trajectory. After growing steadily, from six before 2012 to 17 (2013-2016) and 26 (2017-2020), they dropped to only 15 studies in the most recent period. This modest retraction may reflect pandemic-related disruptions in in-person research, changes in funding priorities, or a focus on scaling previously validated systems rather than developing new interventions. However, the strong technical output suggests that the field continues to innovate at the system level, though perhaps at the expense of new clinical validation.

The distribution of research across these two domains suggests that, despite more than two decades of progress in robot-assisted therapy for ASD, the field remains largely siloed. This raises a important question: Are we building technically impressive systems that lack clinical relevance? Or are we conducting clinically meaningful interventions with outdated or fragile platforms? In the recent years, however, there

[5] A few years displayed a more balanced disciplinary distribution, specifically 2012 (8 technical, 2 clinical, 6 other), 2013 (8 technical, 6 clinical, 2 other), 2019 (18 technical, 10 clinical, 5 other), and 2020 (15 CS, 8 clinical, 3 other). These inflection points likely reflect moments of greater cross-domain collaboration and editorial focus on interdisciplinary themes. This suggests that while disciplinary silos remain, there is a precedent for convergence.

is an evidently growing degree of cross-pollination between technical innovation and clinical application. Although several efforts have directly contributed to this, such as interdisciplinary collaborations, co-authored publications, and special issues, there remains no unified framework for conducting SAR research in a way that fully satisfies both engineering and clinical expectations.[6]

3.3 Intervention Design & Goals

As noted in Section 3.1, the field encompasses a wide range of studies with varying methodologies and objectives. However, this body of work can be broadly categorized into three interrelated yet distinct phases: intervention design, robot development, and evaluation. This section focuses on the first phase: designing the intervention's goals and structure. We begin by summarizing traditional clinical approaches to ASD therapy and then examine how roboticists have adapted these established approaches to create robot-assisted interventions. We outline the social, cognitive, and behavioral objectives these interventions aim to address and we situate these goals within the broader context of age-related needs and developmental trajectories. Finally, we identify gaps in existing approaches and highlight opportunities for more targeted and developmentally appropriate intervention design.

3.3.1 Clinical Foundations of Autism Therapy

Typical interventions for ASD are structured programs designed to teach young children [239, 240] how to initiate, sustain, and respond appropriately in social interactions. These interventions are grounded in well-established behavioral and developmental principles and often follow a highly structured format [241, 242]. Complex skills, such as engaging appropriately with peers, are broken down into low-level, teachable components (e.g., maintaining eye contact, turn-taking, greeting others, or interpreting facial expressions) and are taught explicitly through modeling, role play, and repetition [243–245]. Visual supports, such as social stories and scripted dialogues [246], are commonly used to increase predictability and offer reliable strategies for navigating unfamiliar social situations.

[6]The significant heterogeneity within the autism population supports the use of rigorous single-case experimental designs. It is a methodological error to rely solely on randomized controlled trials as the gold standard for evidence, as they often obscure important individual differences. Likewise, rejecting studies with small sample sizes or rich qualitative data overlooks valuable insights that are critical to understanding early feasibility outcomes.

Generally, humans deduce the unwritten rules of social interaction through everyday observation and exposure [247]. For instance, we learn the social constructs of when to initiate a handshake versus a wave, how close to stand during a conversation, or when to make eye contact and when to look away—all by observing and imitating others around us. These behaviors are rarely taught explicitly; instead, they are absorbed through repeated exposure to contexts where timing, appropriateness, and cultural nuance often defy simple, rule-based explanations. However, for ASD, this kind of incidental social learning may be less accessible. Many individuals with ASD experience difficulties with imitation or interpreting social cues [248], and may not naturally seek out or attend to the social interactions by which these norms are typically learned [249].

Rather than relying on passive exposure, many programs actively shape behavior through reinforcement-based methods, such as Applied Behavior Analysis (ABA), where targeted behaviors are systematically reinforced and other behaviors are either not reinforced or redirected [244].[7] Organic sources of reinforcement appear in peer-mediated strategies [252, 253], in which children practice skills alongside neurotypical peers in small-group settings, allowing for real-time feedback and social modeling.

Crucially, most interventions involve parents or caregivers to ensure that the progress made during structured sessions is reinforced beyond scheduled therapy times or designated environments. The ultimate goal is not merely to teach isolated, low-level social skills, but to foster the capacity for socially meaningful and developmentally appropriate engagement that can generalize beyond the confines of formal intervention. For instance, even a modest improvement in joint attention—the ability to coordinate attention with another person toward a shared object or event—can serve as a critical scaffold for more complex skills such as initiating conversations, interpreting others' intentions, and participating in cooperative play across diverse, real-world contexts.

3.3.2 Targeted Behaviors for Robot Therapy

Roboticists have applied many of these established clinical practices to create SARs that model, prompt, and reinforce key social behaviors in structured and repeatable

[7]Although ABA remains one of the most widely used interventions, it is also highly contested: early approaches often relied on extrinsic rewards (e.g., candy) and repetitive drills (Discrete Trial Training), and at times included aversive techniques that have since been discredited. Contemporary practices, by contrast, emphasize more naturalistic, socially embedded reinforcement strategies. For recent discussions about what ABA is and is not, see [250, 251].

ways. To this end, robots have been developed to support a variety of interaction goals, including capturing and sustaining attention, eliciting joint attention, modeling the expression of empathy, and mediating turn-taking. We outline the specific behaviors that have been the focus of interventions across the studies in our review corpus. While the seminal reviews in 2012 demonstrate the field's exclusive focus on children at that time, the recent decade has seen a notable broadening in user demographics. We can now examine why and in what specific ways the behaviors that are prioritized for early childhood differ from those emphasized in programs for teenagers, for example.

Accordingly, we organize our analysis by age group, based on the predominant participant demographic in studies targeting each specific skill, in order to examine how the goals of robot-assisted ASD therapy align with developmental needs and real-world life contexts.[8] Throughout, we cite representative and influential studies to characterize each developmental stage.

While chronological age serves as a useful lens for examining developmental priorities, it is important to recognize that clinical interventions are also shaped by other participant characteristics such as symptom severity, verbal ability, and cognitive functioning. These dimensions often guide participant inclusion criteria and influence the selection of target behaviors. Therefore, after presenting our age-based analysis, we turn to a complementary examination of how the intervention goals vary between studies stratified by functional profile and diagnostic presentation.

Early Childhood: Joint Attention, Imitation, Symbolic Play

Interventions for early childhood (from birth to 5 years) most often focus on foundational social-communication skills that typically begin to develop in infancy but are delayed or atypical in ASD. One core skill is *joint attention*, commonly defined as the capacity to coordinate attention with another person toward an object, event, or person for the purpose of sharing a social experience [258, 259]. It is a fundamental developmental milestone that typically emerges between 8 and 14 months of age. As a social skill, it extends beyond mere eye contact to include gestures (e.g., point-

[8]Individual studies often included a wide distribution of participant ages, in some cases spanning multiple recognized developmental stages, from early childhood through adolescence. For example, Feil-Seifer and Matarić [254] recruited participants ranging in age from 20 months to 12 years. Other studies also included broad age distributions (e.g., [234, 255–257]). As a result, categorization by age range (even as presented in this section) should be interpreted with some flexibility. Furthermore, it is important to note, while many studies reference age or developmental context, none explicitly report statistical comparisons between different age groups (e.g., comparing outcomes in younger versus older children within the study).

ing, showing), gaze following, and other behaviors that signal shared intentionality. Infants learn not only to follow the direction of another's eyes or head but also to interpret the communicative intent behind those cues and to actively initiate shared attentional states. The ability to engage in joint attention underpins later-developing capacities such as language acquisition, social referencing, and theory of mind [260].

Atypical gaze behavior is a core diagnostic feature of ASD and is closely linked to many social and communicative difficulties characteristic of the condition. As such, it is one of the earliest targets for intervention (robotic or otherwise). More than half of the studies in our corpus ($N = 162$, 53.3%) report robotic interactions aimed at developing gaze-related skills, such as making eye contact, gaze following, joint attention.[9] Of these, 84 studies involved participant samples in which the age range primarily constituted of those in early childhood. This distribution suggests that, while gaze is a central focus in early childhood interventions, it remains a priority across later developmental stages.

In many of these studies, children with ASD exhibit spontaneous joint attention behaviors during their interactions with robots—for example, shifting gaze between the robot and an adult, or pointing toward the robot while looking at another person, with the apparent intent of sharing interest or drawing the other's attention to a specific feature. Children with ASD show this behavior in robot-based interventions despite previously displayed tendencies to avoid eye contact or engagement with parents or therapists. Many studies captured such improvements in joint attention by analyzing gaze patterns between the child and the robot during interactions, using tools such as Tobii eye-trackers or frame-by-frame manual coding. These approaches offer fine-grained insights into real-time engagement behaviors but are often limited to the immediate robot–child context. By contrast, only one study to date, by Scassellati et al. in 2018 [3], explicitly evaluated the generalizability of robot-assisted improvements through a clinically validated, clinician-administered probe (i.e., [261]). This structured assessment was conducted at multiple time points: before the robot was introduced (baseline, pre-intervention), during the intervention, and after the robot was removed (post-intervention). Crucially, it measured the child's ability to demonstrate learned joint attention skills from the SAR intervention in a new social context (with a clinician and without the robot).

Beyond the ability to make, sustain, follow, and share gaze, the ability to then *imitate* behavior is a core mechanism for learning socially appropriate behavior. For instance, young children learn how to greet others, use everyday objects, express emo-

[9] This excludes studies focused solely on general task-oriented attention.

tions, and navigate social routines by observing and imitating the actions of those around them. 87 studies in our corpus (28.6%) target imitation skills. There are several hypotheses as to why imitation is particularly difficult for children with ASD, ranging from reduced mirror neuron activation that affects one's ability to map observed actions onto their own motor systems,[10] to reduced motivation to attend to or engage with others, which limits the opportunities and drive to imitate. Imitation arises naturally in human-robot interaction, as children are encouraged by adults or by the robot itself to imitate the robot's actions (e.g., [209, 266, 267]). Other forms of imitation emerge spontaneously and develop with continued exposure to the robot's behavior. This outcome is often anticipated (e.g., [3, 268]) or results from intentional design choices, as SARs are frequently programmed with exaggerated, salient, or reinforcing behaviors specifically to capture children's attention and encourage engagement (e.g., [269–274]).

Symbolic play is the ability to use objects, actions, or ideas to represent other objects, actions, or ideas during play. It requires the ability to share attention with others (joint attention), observe and imitate modeled behavior, and flexibly coordinate actions within a shared imaginative frame. For example, a child might pretend that a banana is a telephone, or that a block is a car. This type of pretend play typically begins to emerge in neurotypical children around 18–24 months of age and serves as a foundation for more complex behaviors such as narrative construction, perspective-taking, flexible problem solving, and early forms of theory of mind. Children with ASD often show delays or reduced engagement in symbolic play, favoring more literal or repetitive behaviors (e.g., lining up cars instead of pretending they're racing). Many studies in our corpus use pretend scenarios as an interaction structure or delivery method (as further discussed in Section 3.3.3), but rarely as the target skill for intervention. For example, robots may enact emotional narratives (e.g., a sad robot that needs cheering up [233]) or participate in everyday routines (e.g., make-believe grocery store visits [29]), allowing users to practice common social situations. Although many researchers leverage role-play to create engaging narrative environments for human-robot interaction, only two studies explicitly target the development

[10] One may argue that interventions aimed at improving fine motor skills, such as handwriting practice or buttoning a shirt, are not inherently *social* in nature. However, in the context of ASD therapy, motor imitation skills cannot be meaningfully separated from social imitation, as both rely on shared attention, observation, and modeled behavior. Some children with ASD have difficulties with motor coordination, timing, or sequencing, making it harder to physically reproduce observed behaviors, even when the intent is present. Sixteen studies in our corpus targeted motor imitation, also referred to as praxis training. These interventions spanned a wide age range, from early childhood (e.g., [262–264]) to adolescence [265]

of symbolic competence [275, 276].

Middle Childhood: Turn-Taking, Language, and Peer Engagement

For middle childhood (approximately ages 6–12), robot-assisted interventions continue to target foundational skills such as joint attention and imitation, but increasingly integrate these into more complex social tasks relevant to school and peer interactions. Many children with ASD in this age range, even those who are verbally fluent, struggle with pragmatic communication (like taking turns in conversation, asking questions of others, or narrating events) as well as with understanding emotions and others' perspectives. Notably, when comparing the most common behavioral targets across age groups, SAR interventions in early childhood tend to emphasize gaze-related skills, whereas those in middle childhood shift toward language-related outcomes.

Turn-taking is an inherent feature of SAR-based interventions, whether or not it is the explicit target of the design. At its core, human–robot interaction is reciprocal: the robot is programmed to elicit social behavior from the user, and the user, in turn, is expected to attend and respond to the robot's actions. Given that it is embedded as a structural necessity for robot interaction, turn-taking appears as a targeted and measured outcome in 40 studies in our corpus (13.2%). These studies feature collaborative activities (e.g., [277–279]), where participants must alternate roles or contributions; to basic conversational practice, where initiating, responding, and pausing all rely on reciprocal timing; to information-sharing and self-disclosure (e.g., [280]), which require knowing when to speak and when to listen; and to spontaneous help-seeking (e.g., [281]), where the child must recognize an appropriate moment to interrupt or request assistance.

Expressive and receptive language goals at this age often involve expanding vocabulary, improving sentence use, or practicing the back-and-forth of conversation. SARs are used in this capacity to model clear, consistent speech patterns, provide contingent social responses, and create opportunities for low-pressure conversational practice. 65 studies in our corpus (21.4%) support these goals through robot-facilitated dialogue, guided interaction, and social storytelling. For instance, Kim et al. [238] found that children with ASD produced significantly more speech in robot-mediated sessions compared to adult-led ones, suggesting the robot's potential as an embedded reinforcer of verbal output. Pioggia et al. [234] reported improvements in social communication following repeated interactions with an android-based system, while Lee et al. [282] demonstrated that animated robot features outperformed human partners in stimulating communicative responses. These studies often use predictable, engaging routines to elicit both expressive (e.g., labeling, requesting) and receptive (e.g., following instructions) behaviors, reinforcing pragmatic language skills like turn-taking, topic maintenance, and timing.

Together, these skills contribute to more effective peer engagement. As children become better able to initiate and sustain conversations, interpret social cues, and respond contingently to others, they are more likely to be included in play, develop reciprocal friendships, and access peer-mediated learning opportunities. Despite this, our data does not show an expected proportional shift toward triadic or group-based designs in middle childhood (53% compared to 59% of early childhood studies). We discuss this further in Section 3.3.3.

Adulthood: Emotion Expression and Vocational Readiness

Of all behavioral targets of SARs for ASD, *emotion recognition and expression* span a wider range of age groups than all other targeted skills, appearing in nearly equal proportions across early childhood, middle childhood, and adolescence (ages 13–17). In early childhood, emotion recognition is often taught through visual attention to facial expressions or affective cues, frequently paired with gaze and imitation tasks. In middle childhood, interventions shift toward expressive skills such as identifying, labeling, and appropriately responding to one's own or others' emotions, often through dialogue and storytelling. By adolescence (16 studies total, nine of which target emotion-related skills), interventions continue to emphasize recognizing emotions in others and expressing appropriate emotional responses. Notably, there are currently no studies in the corpus that explicitly teach emotional regulation strategies—that is, what to *do* when experiencing stress or anxiety. While a few SAR-based interventions guide adolescent to adults users in recognizing and labeling their own emotional states [283–285], none provide concrete coping techniques or scaffold behavioral responses to help manage those emotions in real time.

Despite growing recognition of the unique needs of adolescents with ASD, work in this age group remains relatively rare in our corpus, comprising only 5.3% of studies. This gap is especially notable given the increasing social and functional demands placed on teenagers as they transition into adulthood. While early and middle childhood interventions tend to emphasize foundational communication and social engagement skills, very few SAR-based studies extend into domains that become salient in adolescence—such as emotional regulation, independent living, or vocational readiness.

Even fewer studies include adults with ASD (ages 18 and up; $N = 9$, representing just 3.0% of our overall corpus), and only four focusing on outcomes relevant to adult social contexts, although exclusively on job interview practice [286–289]. While this reflects an important area of need (supporting employment readiness), the narrow emphasis on interview scenarios overlooks a broader range of challenges faced by adults with ASD. This gap is further surprising given that the first generation diagnosed under broadened diagnostic criteria is now entering midlife, creating an urgent need for developmentally appropriate, scalable supports that extend beyond childhood. Daily living skills, workplace social dynamics, financial literacy, and community navigation remain largely unaddressed in the current SAR literature. Only one study [29] begins to address this broader scope by targeting *interruptions resiliency*, a newly proposed but valued skill linked to both employability and everyday emotional regulation.[11]

Participant Profiles and Functional Variation

While developmental stage provides a useful organizing framework for understanding the goals of robot-assisted autism interventions, it does not fully account for the diversity of needs and capabilities among individuals on the spectrum. In clinical contexts, interventions are often tailored not only to age but also to functional characteristics such as symptom severity, mental age or cognitive functioning, and co-occurring conditions. These variables shape both the design and delivery of therapy and are frequently used to define inclusion criteria in research studies. For instance, in our review corpus, we find that many studies explicitly or implicitly constrain participation to individuals within a specific range of functioning. These functional distinctions often map onto different therapeutic goals, even within the same chronological age group.

The majority of studies in our corpus focus on participants with fewer support needs, sometimes described as "high-functioning".[12] This is likely due to the practical demands of engaging with the study materials: participants are typically required to follow prompts from robots or adult facilitators, attend to structured tasks, and complete standardized assessments. These requirements make it easier to recruit

[11] As this was our study [59], it is described in detail as Chapter 6 of this dissertation.

[12] In this dissertation, we retain the term "high-functioning" when referring to participants, as it remains widely used in many of the studies reviewed here. We acknowledge, however, that this terminology is contested and often critiqued as reductive; we include alternatives such as "individuals with fewer or greater support needs" to better reflect the diversity of profiles [290].

and work with individuals who can readily provide assent and comply with study procedures, but they also systematically exclude those with higher support needs, such as minimally verbal individuals or those with co-occurring intellectual disabilities.

This bias is evident not only in the participant profiles but also in how researchers report diagnostic and inclusion criteria. While all studies report an autism diagnosis (as required by our review screening process described in Section 3.1), the specificity of that diagnosis varies widely. Some papers simply state participants have an ASD diagnosis without indicating the diagnostic tools used, while others include more detailed criteria such as confirmation via ADOS-2 [291], SCQ [292], or clinical judgment. Inclusion criteria often implicitly or explicitly favor participants with fewer support needs—for instance, requiring verbal fluency, typical vision or hearing, or the ability to remain seated and attend to a robot for a set period of time. Although these restrictions are often necessary for experimental control, they are also practical necessities tied to safety, protocol feasibility, and the suitability of intervention targets. These requirements, however, inevitably limit the generalizability of findings and contribute to the ongoing underrepresentation of individuals with more complex communication or cognitive profiles.

Nonetheless, a small number of studies in our corpus explicitly aim to include participants with higher support needs, often driven by a desire to expand accessibility and address therapeutic gaps. For instance, Abu-Amara et al. [293] designed a robot-based intervention for children with moderate to severe ASD to support foundational academic and behavioral goals—recognizing that the lack of research involving this population has left many existing tools ill-suited for individuals with limited verbal or cognitive functioning. Giannopulu et al. [294] worked with children diagnosed with severe autism (CARS score = 43),[13] employing a sensory-integrative framework that used rhythmic movement and sensory feedback to reduce anxiety. Rather than modeling normative behavior, the robot's role was to meet the child at their sensory and behavioral baseline; the authors emphasized the need for interaction strategies grounded in the user's experience rather than neurotypical expectations. Clabaugh et al. [295] included children with mild to moderate autism and emphasized the need for long-duration, personalized adaptation as a way to support children who may

[13]The Childhood Autism Rating Scale (CARS) is a clinician-rated tool used to assess the severity of autism symptoms, with scores ranging from 15 to 60 and scores above 37 indicating severe symptomatology. A score of 43, as reported by Giannopulu et al. [294], reflects a high degree of autistic behaviors in multiple domains. Importantly, such scores specifically index the severity of autism symptoms and should not be conflated with cognitive impairment, which is measured separately (e.g., through IQ or adaptive functioning assessments).

not engage reliably with short, lab-based sessions. Even though the study's primary sample was not composed of children with high support needs, the researchers noted that adaptability would be essential for effectively extending robot use to this group. Chung et al. [296] included children with formal clinical ASD diagnoses but excluded those with severe intellectual disability—indicating an attempt to work with a broader but still bounded range of functional abilities. These examples remain the exception rather than the norm.

Moreover, because most studies limit participation to relatively high-functioning individuals, the field has yet to produce robust within-sample comparisons across different functional profiles. For example, no study has reported stratified outcomes based on the severity of the symptoms (e.g., Level 1 versus Level 3 ASD) or verbal ability. As a result, it is difficult to assess how specific behavioral targets may be more or less effective depending on the user's individual capabilities, or to identify clear strategies for adapting robot behavior to accommodate more complex communication or cognitive profiles.

3.3.3 Structure of Robot Therapy

Beyond the specific behavioral skill targeted, we now consider the structural elements of robot-based therapy as these profoundly shape the design, goals, and outcomes of each study. These elements include the physical setting in which the therapy takes place, the configuration of interactions (e.g., dyadic between just the robot and participants, or small group in several peers interact with the robot alongside the participant), the amount of exposure to the therapy and robot, and the degree of therapeutic contingency built into the therapy design (e.g., activities with the robot are adjusted to a child's progress).

Intervention Setting

The location where SAR therapy occurs plays a critical role in shaping both the feasibility and the ecological validity of the intervention. Over the past two decades, there has been a notable shift in the deployment settings of SAR-based interventions.

In the early years of SAR research, particularly before 2012, robot-assisted ASD therapy was conducted almost exclusively in highly controlled environments such as university laboratories, research clinics, or specialized autism centers. These studies prioritized experimental control and proof-of-concept demonstrations, often featuring brief, scripted interactions between a child and a custom-built or manually operated robot. Access to participants typically required close collaboration with clinical institutions, and the technical complexity of the robots demanded the presence of skilled operators or researchers during every session. While these controlled settings enabled high internal validity and detailed behavioral coding, they offered limited insight into how robotic interventions might function in naturalistic, real-world environments. Overall, 233 studies in our corpus (76.6%) were conducted in either a lab or clinic.

In the decade following 2012, there was a marked increase in studies deploying social robots in real-world environments, including classrooms, therapeutic day programs, and, to a lesser extent, participants' homes. In our review corpus, 52 studies took place in school settings, while only nine described in-home deployments; of real-world environments, these were the two largest categories. School-based studies often integrated robot-assisted activities into existing classroom routines, enabling researchers to observe peer interaction, teacher facilitation, and the naturalistic generalization of targeted skills within authentic educational contexts. Similarly, home studies placed the robot within the child's daily environment, allowing researchers to examine how intervention effects might extend beyond structured sessions, such as through spontaneous interactions with siblings, caregiver involvement, or behavioral changes observed throughout daily routines.

Of the nine home studies identified, seven involved repeated sessions over the course of a week or more, and eight adopted a dyadic structure involving a single child interacting with a single robot. Only one study involved triadic interactions, directly involving the caregiver in the intervention [3]. Only one study in this category targeted adults with ASD [29]. Only one study included a control condition [297] to directly compare outcomes between interventions with and without the robot. All other studies relied on within-subject or between-subject designs to evaluate participant outcomes at discrete time points during the intervention period.

Furthermore, many of these studies reflect on the feasibility of deploying SAR systems in home environments and the unique challenges posed by the dynamic, uncontrolled nature of the household setting. In such contexts, higher levels of robot autonomy are often required, as it is typically impractical for a therapist or experimenter to be physically present during multi-day or week-long interventions. Compared to school-based interventions—where larger sample sizes of children can experience the robot, and where teachers or researchers can serve as consistent third-party facilitators within a structured routine—in-home deployments offer fewer opportunities for in-situ behavioral observation and social generalization.

Despite this diversification (from highly controlled laboratory and clinical environments to more naturalistic, familiar contexts such as the home), the studies in our corpus remain confined to a single physical context. While such multi-context protocols are common in broader behavioral intervention research (as reviewed by [298]), none of the studies in our corpus explicitly implement structured generalization procedures across different settings. For example, we find no instances where robot-based sessions conducted in classrooms are followed by systematic observation in peer-rich, unstructured contexts such as the school playground or cafeteria. This represents a missed opportunity to evaluate whether the gains supported by SARs translate into spontaneous human–human interactions beyond the physical intervention setting.

Interaction Configuration

The most common set-up involves a single child interacting one on one with a robot (N = 169). This dyadic structure is especially common in early childhood interventions targeting foundational skills such as joint attention, imitation, and turn-taking, as a single stimuli (i.e., the controlled behavior of the robot) isolates and makes salient cues for eliciting the target behavior. The nature of dyadic interactions removed other confounding influences on behavior to examine how precise robot behavior can directly impact user outcomes. However, while useful for early-stage learning, dyadic formats often limit the opportunity for practicing contingent, reciprocal behavior in more complex social environments. Most notably, the social target in dyadic SAR studies is often the robot itself despite the intended outcome of improved human-human interaction. This creates a disconnect between the learning context and the transfer context, a limitation that is rarely addressed directly in the evaluation of SAR interventions.

Triadic designs introduce a third participant, typically a therapist, caregiver, or researcher, who scaffolds or guides the interaction (N = 117). These configurations are more common in interventions where interpretation, modeling, or guided reinforcement is needed. The adult may help translate the robot's behavior, provide emotional support, or model appropriate social responses. While triadic structures introduce greater complexity, they offer richer opportunities for generalization such as seeing how robot-child interactions transfer to child-adult interactions.

Despite the central role of peer interaction in real-world social development, relatively few studies in the corpus employ peer-mediated triads, in which a neurotypical or similarly diagnosed peer engages alongside the target child. The limited adoption of this configuration may reflect logistical challenges in recruiting and matching peers or the added variability of introducing another child into the experimental setup. However, omitting peer partners from the intervention design may limit opportunities for practicing socially contingent behaviors and reduces the likelihood of assessing generalization to typical peer environments.

A small number of studies introduce a different type of triadic structure by incorporating two robots as social agents within the interaction [299–301]. This design allows for the simulation of autonomous social scenarios without direct human involvement. For example, the research by Soleiman et al. [302] highlighted the value of two robots in creating a "fully robotic social environment" to teach emotion recognition skills through observational learning. The children observed the two parrot-like robots engaging in social interactions, specifically discussing facial expressions, which proved effective in improving their emotion recognition capabilities. The study emphasizes that this multi-robot setup is crucial for simulating complex social situations that a single robot cannot provide without human intervention.

In summary, the majority of studies continue to rely on configurations in which the robot is both the instructional agent and the *only* social target. While this simplifies the design and measurement of outcomes, it constrains the types of social behaviors that can be meaningfully practiced. For instance, taking turns with a robot that uses fixed timing cues does not fully capture the nuanced demands of conversational turn-taking with a peer, whose responses are inherently less predictable. In general, the majority of studies do not include follow-up sessions or evaluate outcomes with a human partner, leaving open the question of whether observed gains persist outside the robotic context.

Session Frequency & Duration

Another key structural variable in SAR-based autism interventions is the frequency and duration of intervention sessions, which directly influences the feasibility, therapeutic impact, and user familiarity with the system. Earlier studies in this field (particularly those conducted before 2012) were predominantly structured as single-session experiments or very short-term trials involving 2–3 sessions. These designs were often intended to test the feasibility of robot use, gather initial user responses, or pilot specific behavioral tasks (e.g., joint attention, imitation). While useful for proof-of-concept validation, such short interactions provided limited insight into the sustainability of engagement, learning trajectories over time, or the potential for generalized behavioral change.

Meaningful behavioral change does not occur in a single session spanning a few minutes. Also, children generally do not exhibit predictable or consistent behavior on a daily or even hourly basis. Therapy sessions can feature a child highly engaged and sharing toys with a peer one day, and the same child distracted, angry, and refusing interaction the next. Human therapists are prepared to handle these changes in mood and preferences; SAR systems designed to interact with children must feature the same kind of adaptability if they are to be fully integrated into therapy [303].

Recent years have seen a growing trend toward repeated sessions (more than one session with a robot) and longer-term deployments (for more than one day), reflecting increased confidence in the usability of SARs and a change in research goals towards measuring therapeutic outcomes over time. A few studies ($N = 20$)[14] now feature intervention periods spanning several days to multiple months, with children engaging with the robot across 5–20 sessions or more. These studies enable researchers to evaluate not only immediate behavioral responses but also changes in skill acquisition, novelty and habituation effects, and sustained engagement over time. In response to the growing emphasis on longer-term interventions, researchers have begun to explore how both intervention design and robot behavior can support sustained user engagement. This includes developing strategies for personalization and adaptation that allow the therapy to evolve in alignment with the user's progress over time. We discuss this further in the next section.

Therapeutic Contingency and Adaptation

SAR interventions vary widely in their level of therapeutic contingency—the degree to which the robot adapts its behavior to the user's needs, behaviors, or general states. At one end of the spectrum are fully scripted, non-adaptive systems that follow fixed routines, offering consistency and predictability but no user-specific personalization. Of the 272 studies in our corpus that reported sufficient methodological detail, 97 employed such scripted designs, in which the robot's behavior was identical across participants, regardless of individual differences. These approaches are especially common in early intervention settings. For example, [304] used a fixed-sequence design to model emotional expressions with the KASPAR robot.

Adaptive behaviors may include adjusting the pace of the session based on user engagement, providing different types of prompts depending on user response, or

[14] These twenty studies are examined in greater depth in Chapter 2. There, we contextualize their features and findings within the broader literature on long-term human–robot interaction, extending beyond the scope of ASD-focused therapy.

recognizing and responding to emotional cues in real time. Some systems allow for therapist-in-the-loop adaptation, where a human operator modulates the robot's behavior based on observation or clinician judgment. Others are more fully autonomous, adjusting parameters algorithmically in response to sensed input. A few recent systems employ reinforcement learning or supervised adaptation mechanisms to tailor the interaction over time based on user progress or preferences.

At the other end are highly adaptive systems that use multimodal sensing (e.g., gaze tracking, voice tone, facial expressions) and behavioral models to personalize timing, content, and expressive output. These behaviors may manifest as adjusting the pace of the session based on user engagement, providing different types of prompts depending on user response, or recognizing and responding to emotional cues in real time. For instance, Lemaignan et al. [284] present a robot modulating its emotional responses based on children's detected anxiety levels, while Clabaugh et al. [295] present a robot that adapts game difficulty based on child's task performance in order to find the optimal challenge level for each child. Some systems introduce therapist-in-the-loop adaptation, where a human operator modulates the robot's behavior based on observation or clinician judgment, such as in [272, 305], which combined therapist oversight with affective modeling to deliver more nuanced therapy.

Human-human social behavior is inherently nuanced and complex, and requires adaptation and flexibility on the spot. Most adaptive robot systems, even if they adjust to the user, do so only in a limited, predefined way. They tend to personalize just a few specific parameters (like timing, prompt frequency, or choice of feedback) rather than adapting in a deep or generalizable way across different users or contexts. For example, we have yet to see a robot that can respond relevantly and contingently to a user's spoken input in a way that resembles natural conversation. Most systems rely on keyword detection or predefined dialogue templates, and cannot engage in spontaneous, improvised exchanges. A truly adaptive system would understand not just the words but the intent and context of a child's statement (e.g., "I'm tired of this game") and respond in a way that meaningfully advances the interaction (e.g., "Let's take a break or find something new. Do you want to instead [...]?"), while preserving therapeutic goals.

Another form of adapation that is currently absent from any existing SAR intervention is adjusting the overall structure of therapy based on the user's evolving needs across sessions. For instance, rather than remaining fixed in a dyadic configuration, the robot could detect the presence of a sibling and initiate a transition to a shared interaction ("Would you like to play this together with your brother?").

Alternatively, if the child initiates an interaction with a nearby human (e.g., looking toward a parent or sharing a toy), the robot could adapt by reducing its own activity or by facilitating that human–human interaction. No current SAR system for autism therapy demonstrates this kind of structural flexibility where the robot can detect and act upon real-world social dynamics that extend beyond itself.

More broadly, while some SAR systems attempt to personalize content or interaction based on user behavior (e.g., changing the robot's prompts based on gaze direction or emotional expression), the rationale for why a specific adaptation cue should lead to improved outcomes is often underexplored. There is frequently an implicit assumption that detecting a behavior (e.g., a frown or looking away) and responding with a fixed strategy (e.g., encouragement or repetition) will improve engagement or learning, but these mappings are rarely grounded in theory or empirically validated. A more rigorous approach would link adaptation strategies to mechanistic models of learning, emotion regulation, or social motivation—allowing researchers to test not only whether personalization improves outcomes, but why and for whom it is effective.

Lastly, feedback is a valuable element of any behavioral intervention. SARs typically deliver feedback in one of two primary forms: *implicit feedback*, where the robot models behaviors or facilitates repeated practice without directly commenting on performance; and *explicit feedback*, where the robot overtly indicates whether an action or response is correct or incorrect (e.g., verbal praise, correction, or reinforcement cues). Most SAR interventions do not provide direct feedback on the quality of the user's social behaviors (e.g., "Please look at me next time." or "I noticed you hesitated to..."). Instead, feedback is typically tied to external task performance, such as completing a level in a game, earning points, or progressing through a predefined activity sequence. While such task-based rewards can motivate engagement, they often fail to explicitly reinforce the specific social skills the intervention intends to develop. More targeted social feedback could be valuable for helping users link their actions to the underlying social goals of the interaction. At the same time, implicit feedback can vary in strength—from subtle modeling of behaviors to more structured forms of guided practice—offering different pathways for reinforcing target skills without overt correction.

Introduction and Exit Strategies

The structure of SAR therapy is not only shaped by where and how sessions occur, but also by how the therapy is introduced to users (also referred to as "onboarding") and concluded ("offboarding"). The introduction sets expectations, frames the robot's role, and can significantly influence early user engagement, trust, and willingness to participate. The conclusion, in turn, plays a critical role in facilitating emotional closure, reinforcing what was learned, and shaping how users reflect on the overall experience. Particularly in child-focused interventions, well-designed exit strategies, such as gradual offboarding, structured goodbyes, or reflective activities, can help mitigate confusion or distress when the robot is removed at the end of the study. Despite their importance, onboarding and offboarding processes are often underreported or treated as peripheral to the intervention.

In early SAR studies, onboarding was often implied or minimal. Participants were brought into a lab or clinical setting, the robot was introduced as part of the experimental procedure, and children were expected to engage without structured warm-up. However, as interventions moved beyond the lab setting into homes and other real-world spaces, many studies began to integrate intentional onboarding practices. These include preliminary sessions where the child simply gets acquainted with the robot, introductory explanations of the robot's purpose using age-appropriate language [306], and initial games or gestures designed to build rapport. For example, Scassellati et al. [3] describe a month-long in-home deployment where the robot was introduced with a narrative backstory: it had crash-landed on Earth and needed help rebuilding its rocket. This story framed the therapeutic activities, games in which the child and caregiver took turns helping repair the robot's rocket.

In contrast to the attention given to onboarding and engagement strategies, relatively few studies describe deliberate offboarding procedures or exit strategies. This omission may stem from the short duration of many interventions, a focus on technical validation rather than emotional continuity, or assumptions that users will naturally disengage once the robot is removed. However, especially in longer-term or home-based deployments, the absence of structured closure can lead to confusion or emotional distress, particularly for younger users who may quickly form attachments to the robot.

For example, rather than abruptly removing the robot, signaling the end of the intervention several days in advance can allow users to anticipate the end and emotionally adjust. The robot itself can participate in this preparation by introducing short reminders, such as noting how many sessions remain. Incorporating a structured exit ritual during the final session can further support closure. This may include the robot summarizing shared activities, prompting the user to reflect on progress, or highlighting favorite interactions from the deployment. To reinforce the value of the experience, providing users with tangible takeaways (e.g., printed summaries of skills practiced or personalized notes acknowledging their growth) can serve as a meaningful marker of achievement and offer continuity after the system is removed.

How the robot is introduced at the start of the intervention or removed at its conclusion is often overlooked in the literature, as these details are rarely reported or explicitly considered in study design. Still, thoughtful onboarding and offboarding strategies can play a valuable role in contextualizing and motivating the intervention. Considering these elements supports a more holistic understanding of the intervention pipeline and highlights important design choices that are typically overlooked.

3.4 Design of Robot Form & Function

At its core, robots are meant to be a stimulus for eliciting desired behavior in users. To achieve this, roboticists have applied established clinical practices (summarized in Section 3.3.1) to create SARs that model, prompt, and reinforce key social behaviors in structured and repeatable ways (overviewed in Section 3.3.2).

More than simply putting a psychosocial theory onto an embodied robot platform, SAR research encounters new questions about how to create acceptable, intuitive, user-friendly systems. For instance, what are the circumstances in which people accept an assistive robot in their environment (e.g., [307–309])? How can we model the behavior of and encouragement by the robot therapist as a function of user personality or cognitive profiles (e.g., [310, 311])? And how can SARs initiate interactions in ways that feel immediately approachable and socially meaningful, without requiring extensive training or acclimation (e.g., [312, 313])? In the previous section, we examined how larger questions related to interaction setting, configuration, and personalization have been addressed through the design of the overall intervention. In this section, we shift focus to how the robot itself is situated within the broader therapeutic framework.

A person engaging in robot-based therapy first notices the physical appearance of the robot. As the therapy session unfolds, the person can then form their understanding of the robot's behavior. Researchers treat these moments as foundational to system design, focusing on two core questions: (1) how should the robot *look*? and

(2) how should the robot *behave*?

3.4.1 Form: How Should the Robot Look?

The physical appearance and form of a SAR is a non-trivial design choice. Prior research has explored a spectrum of embodiments—from highly anthropomorphic robots with silicone-based skin and expressive facial musculature (e.g., KASPAR [304], Actroid-F [314], Milo [315]), to visually simplified systems that are cartoon-like (with oversized and exaggerated primary features such as eyes or exempt of secondary features like eyebrows or lower eyelids; e.g., Sota and CommU [316], NAO [228], QTrobot [229]), to low-actuation, limbless or stationary, animal-like systems (e.g., Paro [172], Dragonbot [150], Pleo [317], Keepon [85]). Table 3.1 summarizes characteristics of ten robot platforms frequently used in the studies included in this review.

These aesthetic decisions are rarely arbitrary and instead reflect competing priorities in therapeutic design: on one hand, the desire to create an engaging, relatable partner that can adequately convey social cues; on the other, the need to avoid overstimulation, uncanny valley effects, or distraction. This is a particularly important consideration when designing for users with heightened sensory sensitivity or social anxiety. We outline a few nuanced aspects of robot form below, focusing on specific design considerations that arise in the context of ASD therapy.

Table 3.1: Summary of Robot Platforms Commonly Used in Autism Therapy Research. This table summarizes commonly used social robots in ASD interventions, highlighting each platform's physical dimensions, degrees of freedom (DoF), sensing capabilities, expressive features, and mobility type. Platforms vary in embodiment and interactivity, reflecting the diversity of hardware employed in SAR research for ASD therapy.

Robot	Type	Height	Weight	DoF	Key Sensors	Expressive Features	Mobility
NAO [228]	Humanoid	58 cm	4.3–5 kg	21–25	Cameras, microphones, tactile sensors, sonar, IMU	Joint movement, LEDs, text-to-speech	Bipedal
KASPAR [304]	Humanoid	55 cm	15 kg	17	Cameras, touch sensors	Minimal expressiveness via silicone face	Static (remote-controlled)
QTrobot [229]	Humanoid	63 cm	5 kg	≤ 10	Cameras, emotion recognition system	Screen-based facial expressions, LED lights	Tabletop (static)
Pepper [173]	Humanoid	1.2 m	28 kg	17–20	Touch sensors, cameras, IR, bumpers, IMU	Tablet interface, speech, LEDs	Wheeled
Pleo [317]	Zoomorphic (Dinosaur)	18 cm	1.5 kg	10–15	Light sensors, tactile sensors, camera	Animatronic movement, vocalizations	Quadruped walking
Keepon [318]	Non-humanoid	25 cm	1 kg	1–2	Microphone, accelerometer	Bouncing, head tilt, eye movement	Static with bounce
Robota [319]	Doll-like humanoid	45 cm	5 kg	10	Tactile sensors, microphone	Head and arm gestures, postures	Static (sitting/standing)
Zeno / Milo [315]	Humanoid	63 cm	6.5 kg	36	Stereo cameras, microphones, touch	Frubber face, natural gestures, speech	Wheeled or walking
Cozmo [320]	Toy-like	15 cm	0.3 kg	5–6	Camera, IR sensors	Screen face, animated expressions, arm gestures	Tracked/wheeled

Note. IMU = Inertial Measurement Unit; IR = Infrared; kg = kilograms; cm = centimeters; m = meters.

Material Properties

The choice of physical materials is a critical design consideration for social robots in ASD therapy, particularly for encouraging tactile interaction. Examples include the soft, durable Veltex exterior of a fish-shaped robot designed for non-verbal children with autism [321], the silicone skin used for KASPAR [304], and the rubberized, squeezable body of Keepon, whose tactile design was intentionally created to encourage physical interaction and withstand repetitive touch [85].

The integration of touch-responsive epidermal coverings, such as RoboSkin on KASPAR [304], and the pressure-sensitive foam layers used in the Huggable robot [322], allows robots to detect and appropriately respond to various forms of touch. This capability is vital for teaching appropriate force modulation during physical contact and for conveying emotional responses through haptic cues. The use of soft, touch-responsive materials and haptic feedback highlights that tactile interaction is not merely supplementary, but can serve as a meaningful communication channel for children with ASD—especially when verbal or nonverbal expression is limited.

While recent work emphasizes the use of soft, textured materials and haptic feedback in SARs, the importance of tactile design is not new. Some of the earliest socially assistive robots were intentionally designed with soft, compliant bodies to invite physical interaction. The broader robotics literature beyond just SARs for ASD therapy may commonly imagine robots to be cold, hard plastic devices that rely predominantly on visual or auditory communication. However, there is a marked difference in design choices when designing for users with ASD, many of whom experience sensory integration challenges. We observe an increasing use of diverse materials to promote tactile interaction, the design of more holistic, multisensory experiences, and the development of systems that leverage touch as a meaningful mode of therapeutic engagement.

Safety & Mechanical Complexity

Safety in human-robot interaction is a multidimensional design concern that encompasses physical, emotional, and social considerations. When working with children with ASD, who may exhibit sensory sensitivities, motor planning difficulties, or unpredictable responses to novel stimuli, the safety of the robot is critical. To mitigate physical risk, researchers frequently adopt conservative design principles, such as limiting a robot's movement speed, range of motion, and applied force. Many robotic platforms used in ASD interventions also incorporate safety-oriented structural features—for example, avoiding sharp edges, covering joints to reduce the risk of pinching, and using compliant or soft materials in areas where contact may occur. The Pepper robot, for instance, includes soft padding in its arms and compliant joints to help minimize physical harm during interaction.

Robots with a high number of degrees of freedom (DoF) can produce more lifelike, fluid, and varied movements, which can support naturalistic gestures and rich social cues. However, this added complexity often comes with significant tradeoffs in cost, reliability, and durability—especially in studies that require repeated or long-term use in real-world environments where technical support and supervision is limited. Motors can burn out, sensors may drift or degrade, and calibration requirements increase as systems grow more sophisticated. High actuation increases the likelihood of mechanical failure and places greater demands on power and control systems.

For many autism interventions, simplified mechanical designs are intentionally chosen to limit the range of possible behaviors and to reduce the cognitive and sensory demands placed on the user. For example, Robota uses simple head and arm gestures to support imitation games [319], while Keepon expresses affect using just four motors for bouncing, tilting, and turning [270]. These streamlined behaviors can still convey attention and emotion effectively, without overwhelming children or risking hardware failures.

Mobility

Researchers must make strategic decisions about a robot's level of physical expressiveness and mobility based on the goals of the interaction, technical feasibility, and the demands of the therapeutic setting. As most SAR interventions are designed for children, and given that children are the demographic least prone to staying in one place for a therapy session, the robot's mobility becomes another crucial factor. While some robots are designed to move freely through space (rolling or walking alongside users), many remain stationary on tables or stands. Fixed robots offer greater mechanical stability, reduced safety risks, and tighter control over interaction context, making them well suited for structured therapeutic tasks. In contrast, mobile robots allow for dynamic, spatial interactions such as following a child, guiding navigation, or exploring a shared environment. These interactions can be more engaging and naturalistic, but also introduce new technical challenges in autonomous navigation and behavior coordination.

Cost and Maintainability

The cost and maintainability of a robot are critical practical concerns that shape platform selection in SAR research. Many of the robots used are commercial-grade platforms with price tags ranging from $\$10,000$ to over $\$30,000$ USD. In practice, costs extend well beyond this initial purchase. Additional expenses typically include proprietary software licenses, developer toolkits, maintenance contracts, accessories (e.g., docking stations, sensors, carrying cases), and replacement parts such as motors, servos, or skins that wear down with use.

Moreover, commercial platforms may impose constraints on repairability and customization. Some are closed systems that restrict access to internal hardware or require manufacturer servicing for even minor malfunctions. This introduces delays in the study timeline, limit on-site troubleshooting, and increases long-term dependence on external vendors.

For smaller research labs and institutions, these challenges can significantly constrain the scalability and longevity of a study. On one hand, deploying a single robot may be feasible for a tightly controlled pilot study. On the other hand, scaling up to five or 10 units for multi-site trials or parallel user testing can dramatically increase financial, technical, and logistical demands. Each additional robot multiplies the burden of setup, calibration, software updates, repairs, and storage. These demands can fundamentally shape the nature of a study: researchers may be forced to stagger participant sessions rather than run them concurrently, reduce sample sizes, or reassess the feasibility for long-term deployment.

As a result, many teams must weigh the expressive and technical affordances of high-end commercial robots against the economic and logistical realities of sustained research use. In some cases, these constraints motivate the adoption of more minimalist platforms (such as tablet-based robots or custom-built devices using off-the-shelf components) that offer limited but sufficient functionality at a fraction of the cost.

3.4.2 Function: How Should the Robot Behave?

The behavioral design of SARs (how they speak, move, respond, and guide interaction) has prompted rich exploration. Some systems emulate the structured, directive style of a clinician or interventionist (e.g., prompting a child to make eye contact or initiate turn-taking; [323, 324]), while others adopt more peer-like or playful interaction styles aimed at encouraging spontaneous engagement (e.g., storytelling, imitation games; [318, 325]). We organize these considerations of the *role* a robot should function within the therapy context, and how SAR research has approached behavioral *autonomy* and *adaptation*.

Roles of the Robot

In robot-assisted autism therapy, the role a robot assumes shapes not only its behavioral repertoire but also how it is perceived and engaged with by children. Robots can adopt various roles, even within the same therapy session and depending on the goals of the intervention. A robot may function as an instructor, modeling behaviors, giving prompts, and guiding activities. It may also act as a responsive toy or social mediator, reacting to the child's behavior and facilitating interaction between the child and others. The roles of robots in our corpus can be broadly categorized into three primary types: instructor ($N = 205$, or 67.4% of all studies), peer ($N = 64$, 21.1%), and mediator or facilitator ($N = 17$, 5.6%).[15] We outline each of these roles and the distinct advantages and considerations they entail.

Robots in a **instructor** role are programmed to deliver structured guidance, often mimicking the interactional style of a therapist or educator. These robots provide clear prompts, reinforcement, and correction to teach specific skills such as joint attention, turn-taking, or emotion recognition. This approach is particularly valuable in early interventions where consistency and repetition are essential. For example, humanoid robots like NAO have been used to instruct children in emotion labeling through direct teaching protocols [326]. On one hand, instructor robots offer high consistency in task delivery, can provide unlimited repetition without fatigue, and can reduce social pressure by depersonalizing correction. On the other hand, these robots may be perceived as less socially engaging or natural, and their directive style may limit opportunities for spontaneous interaction or generalization beyond the therapy context.

Robots designed to act as **peers** emphasize playful, bidirectional engagement. These systems typically exhibit more naturalistic behaviors and are often introduced as playmates or companions. Peer-role robots may participate in turn-taking games, mimic the child's actions, or join collaborative storytelling. Robots that adopt a

[15] While these role categories provide a useful framework for organizing the literature, they are not always mutually exclusive. In practice, robots can exhibit characteristics that span multiple roles. For instance, a robot may engage in playful, peer-like exchanges while simultaneously guiding or facilitating triadic interactions with a nearby adult or peer. Nevertheless, for the purposes of analysis, we assigned a single primary role label to each study. This classification was based on the role explicitly stated in the paper, the nature of the robot's behavior (e.g., providing corrective feedback is indicative of an instructor role), and the type of interaction it predominantly supported.

peer role hold several advantages. Peer robots foster social reciprocity and reduce the power imbalance common in adult-child therapy. They are especially effective in eliciting spontaneous social behaviors, including joint attention, smiling, and vocalization. Yet, the unstructured nature of peer-like interactions may be less effective for skill acquisition that requires precise instruction, and some children may struggle to interpret the robot's play cues without adult mediation.

In the **mediator** or **facilitator** role, the robot is positioned as a bridge between the child and other humans. Rather than being the central partner in interaction, the robot helps scaffold or structure joint activities. For example, a robot might prompt the child to ask their parent a question, guide collaborative storytelling between children, or monitor and adjust the pacing of shared tasks [278, 327]. Here, robots as mediators promote triadic interactions and facilitate generalization of social skills to human partners. They can help reduce anxiety associated with direct social interaction by redirecting attention toward a shared focus. However, the robot's success in this role depends on its ability to manage multiparty interaction, which remains technically challenging.

Most interventions in SAR research still feature robots in instructional roles. However, a smaller subset of studies explores a *learning-by-teaching* paradigm, in which the child takes on the role of instructor and the robot as the learner. This dynamic requires the user to reflect, organize, and verbalize their knowledge. Doing so to help teach a robot may allow users to internalize the target behavior more effectively than the comparatively more passive dynamic of receiving instruction. Teaching another (whether a human peer or robot) inherently creates a sense of accomplishment which then promotes intrinsic motivation for learning and self-efficacy. For example, Zaraki et al. [328] designed an intervention in which pairs of children played a "treasure hunt" game by hiding toys in different locations and then teaching a robot how to find them. The robot would attempt to guess which toy was in which location and the children gave explicit feedback—a "Yes, you're right" or "No, that's wrong." Through trial-and-error and the kids' feedback, the robot gradually learned the correct associations between six toys and three locations. To achieve this dynamic, the robot's behavior relied on a simple reinforcement learning algorithm and the ability to express its uncertainty verbally ("Hmm, I'm not sure") during the game; these design choices kept the children informed about why the robot was making mistakes or changing its guesses, so they could adjust their teaching strategy. In an another study, Barnes et al. [329] designed a musical dance game where a child and robot took turns dancing and imitating each other. A tablet was used to structure the activity (displaying dance moves or music), and the robot was programmed to learn the child's dance moves by mimicking them, effectively letting the child "teach" the robot new dance steps. Case studies showed that children were eager to show the robot what to do. The novelty of being the robot's teacher made some typically shy children open up. In some cases, the user became strongly motivated over time to continue teaching the robot, and this process naturally elicited core social behaviors like increased eye contact, gesturing, and even conversation directed at the robot or nearby adults.

Balancing Autonomy and Adaptation

One common interpretation of "autonomy" in human-robot interaction refers to a system's ability to operate within expected parameters without requiring human involvement or supervision. Under this definition, autonomy can span a broad spectrum: from fully scripted behaviors that ignore user input to highly adaptive systems that respond dynamically to the nuanced expressions of humans in their environment. Unsurprisingly, systems on the scripted end of the spectrum are often more likely to meet the criteria for autonomous operation because their behavior is predictable and rule-based. In contrast, systems that rely on interpreting and responding to complex human behavior introduce greater variability and unpredictability, making it more difficult to define and guarantee "expected" outcomes using traditional if–then logic. We categorize the systems in our corpus into three levels of autonomy: fully autonomous ($N = 100$, 33.4%), semi-autonomous ($N = 183$, 61.2%), and non-autonomous ($N = 16$, 5.4%)—based on the 299 studies that provided sufficient detail to determine autonomy level. Each level entails different trade-offs in terms of technical complexity, experimental control, user engagement, and ecological validity.

In **fully autonomous systems**, the robot perceives, processes, and acts without human intervention during the session. These systems often integrate real-time sensing (e.g., cameras, microphones, gaze tracking), behavior recognition, and decision-making algorithms to adjust their actions in response to the child's behavior. For example, a fully autonomous robot might adapt the difficulty of a task based on the child's performance or initiate new behaviors when engagement appears to wane. While these systems offer scalability and reduce the need for human oversight, they face significant challenges in reliability, particularly when deployed in real-world settings with variable lighting, noise, or user behavior. Errors in perception or misinterpretation of user intent can quickly erode trust or disrupt the flow of interaction.

Semi-autonomous systems blend autonomous control with human oversight or intervention. In these setups, the robot may handle certain actions automatically (e.g., delivering prompts or gestures), while a human operator monitors the session or selects among pre-defined behaviors behind the scenes (a paradigm often referred to as Wizard-of-Oz control). This approach is common in early-stage prototypes or exploratory studies, where researchers prioritize flexible interaction and safety over full automation. Semi-autonomy allows for more naturalistic engagement while retaining human judgment to resolve ambiguous situations. However, it also introduces constraints on reproducibility and scalability, and it may mask the system's actual capabilities if users are unaware of the operator's involvement.

At the other end of the spectrum, **non-autonomous systems** are fully teleoperated or scripted in advance. These robots follow a pre-set sequence of actions or are controlled in real time by a human facilitator. Such systems are valuable for isolating specific variables, maintaining experimental consistency, and avoiding technical failures. For instance, in early proof-of-concept studies, researchers often used remote-controlled robots to demonstrate particular behaviors or social cues without needing autonomous perception. While non-autonomous systems lack adaptability, they ensure tight experimental control and can simulate social interaction adequately enough to elicit meaningful responses from children with ASD.

Taken together, these levels of autonomy reflect not only technological capabilities but also design philosophies and research goals. Fully autonomous systems prioritize scalability and ecological validity but risk unpredictability. Semi-autonomous systems balance adaptability with control. Non-autonomous systems prioritize control and consistency but limit personalization. Still, achieving robust, full autonomy remains technically demanding. In the past decade, many systems were deployed in controlled settings such as labs and clinics (as summarized in Section 3.3.3), where it was more feasible to map low-level if–then rules to already constrained participant-robot behaviors. However, recent work has shifted toward deploying robots in naturalistic environments (such as homes, schools, and community centers) where social interactions are less predictable and environmental variability is higher. Robots must now not only respond to user behavior but also contend with variability in its physical environment (e.g., fluctuating lighting, background noise, the presence of multiple people or objects in the scene). This shift has encouraged exploration into adaptive control strategies, real-time personalization methods, operationalization of norms and judgments of behavioral appropriateness.

Even as ASD interventions increasingly move into real-world contexts for longer-term deployments, the design of SARs continues to reflect artificial constraints to ensure control and predictability. This emphasis is justified: tightly scripted interactions and rigid activity flows help ensure safety, experimental replicability, and consistent data collection. However, these design constraints often result in narrowly defined therapy sessions that occur only in designated physical spaces (e.g., seated at a desk to remain in the robot's field of view, even within a home setting) and within fixed "therapy time windows," only initiated by the robot itself. The robot's behavior is largely decoupled from the broader rhythms of everyday life, limiting opportunities for spontaneous or proactive user engagement, real-time adaptation, or seamless integration into everyday routines. In our corpus, this pattern holds across all SAR-directed therapies conducted in participants' homes. We found no study that allows users to independently initiate interactions outside of a pre-scheduled time or freely engage with the robot while going about their daily activities. For example, we have yet to see a SAR that enables a user to spontaneously choose to practice their conversational skills while they are washing dishes or watching television. Instead, existing work remains anchored to static physical positions and predefined schedules, reflecting a robot-first design paradigm rather than a user-centered or context-responsive model.

3.5 Evaluation of Robot Therapy

Measuring the success of SARs in autism therapy remains an evolving and debated area. Early studies typically relied on in-session metrics, such as frequency of eye contact, gesture imitation, or task compliance, as proxies for therapeutic benefit. However, children with ASD are often highly sensitive to novel stimuli and changes in routine [194], making the amount of exposure they have to a robot a critical factor when evaluating the robot's therapeutic effect; a child's initial reaction to the robot as a novel stimulus may differ significantly from their response once the robot becomes familiar. While short-term behavioral measures provide useful insights into immediate engagement and social influence [330, 331], they tend to reflect novelty or exposure effects rather than offer conclusive evidence of long-term developmental gains or generalizability. Very few SAR studies have followed users over extended periods with the scale and methodological rigor necessary to draw conclusions that would be considered robust by clinical standards. This was one of the main critiques in both seminal reviews of 2012. Both Scassellati et al. [20] and Diehl et al. [206] characterized the standard of SAR methodology as having tiny sample sizes, absence of control groups, and minimal follow-up, without clear conclusions about efficacy.

Over the past decade, there has been a concerted effort to address these limitations. As noted in Section 3.2.2, there is growing interdisciplinary collaboration between the robotics and clinical communities, and robotics researchers are now more attuned to clinical priorities and standards than ever before. Begum et al. in 2016 [332] explicitly called out that clinicians were "not convinced" by the early robotics studies because "the vast majority of HRI studies on robot-mediated intervention do not follow any standard research design and consequently the data... is minimally appealing to the clinical community." They urged roboticists to adopt clinical trial guidelines and focus on demonstrating utility, not just likability. The response to this call is evident in the recent decade: more studies now have control conditions, baseline measures, and standardized outcomes, aligning with the evidence-based practice framework that clinicians expect.

Study Design and Sample Sizes

Where earlier work often amounted to anecdotal case studies or proof-of-concept pilots involving less than five participants, we now see substantially larger trials. Our review identified 19 randomized controlled trials (RCTs)—a remarkable increase from essentially zero true RCTs in 2012.[16] These RCTs also include sample sizes that were unheard of a decade ago. For example, van den Berk-Smeekens et al. [333] enrolled 73 children in a 3-arm trial. Zheng et al. [334] conducted an RCT with 20 toddlers with ASD to test a robot-mediated joint attention intervention. Marino et al. [335] ran an RCT with 14 children to evaluate a robot-based socio-emotional skills training. Although these numbers remain modest by clinical trial standards (where sample sizes often range from several dozen to hundreds of participants), they represent a clear improvement over earlier SAR studies, which frequently included only a handful of participants.

In parallel with the rise of RCTs, many recent studies now incorporate control or comparison conditions—features that were often absent in earlier SAR research. Common experimental designs include comparisons between robot-assisted therapy and treatment-as-usual, or between robot-assisted and human-only versions of the same intervention. These designs aim to isolate the contribution of the robot itself. For example, Marino et al. [335] compared a CBT-based emotional intervention delivered

[16] A *randomized controlled trial* is a rigorous experimental design in which participants are randomly assigned to either an intervention group (receiving the robot-assisted therapy) or a control group (receiving standard treatment or no intervention). This methodology minimizes selection bias and enables clearer attribution of observed effects to the intervention itself.

by the humanoid robot NAO versus the same content delivered by a human alone; children in the robot-assisted group demonstrated greater improvements in emotion understanding. Similarly, Zheng et al. [334] evaluated a robot-based joint attention intervention against a no-robot control. Although the group differences were small and not statistically significant, the study nonetheless contributed valuable insights into ASD outcome variability and helped establish a framework for more rigorous evaluation.

Establishing a robust baseline is essential for assessing the impact of any intervention. However, for studies conducted outside of controlled clinical or laboratory environments, implementing equivalent non-robotic control conditions is often challenging. As a result, many studies, particularly those involving small samples or personalized therapies, adopt within-subject designs where each participant serves as their own baseline. This approach is common in ASD research due to the high variability in individual developmental trajectories.

Before questions of *efficacy* can be meaningfully addressed, many robot studies first confront the challenge of demonstrating *feasibility*. Deploying robots in real-world contexts—such as homes, clinics, or classrooms—requires overcoming significant technical barriers, including system autonomy, reliability, and context awareness [225]. Moreover, recruiting participants from such a specialized and highly protected population presents significant challenges, often leading to studies with small sample sizes and limited statistical power. This recruitment difficulty also tends to skew the demographic profile toward individuals considered "high-functioning," who possess the language and functional abilities needed to engage with study materials and protocols. For these reasons, early-stage research often prioritizes evaluating whether a system can function safely, consistently, and acceptably in a given environment, as opposed to whether it produces measurable therapeutic change.

In this early-stage context, conventional notions of evaluation rooted in clinical trial standards may not be the most appropriate or practical. In clinical practice, measuring progress in social skills is often subjective and infrequent, relying heavily on therapist observations or parent reports. In contrast, robots and their associated sensors (cameras, microphones, even wearables) can provide continuous, objective tracking of behaviors. For example, a robot's vision system can count how many times a child initiates eye contact each session, or how their latency to respond changes over time. These metrics can be graphed and used by clinicians to make data-driven decisions (much like a behavior analyst uses frequency counts in ABA). Where robotics brings the potential for more fine-grain detection, analysis, and prediction to the therapeutic context, it introduces an alternate view of "sample size." SAR researchers that aim to evaluate system feasibility would examine system performance over user outcomes. Therefore, "sample size" may refer to the number of system actions or interaction sessions, in contrast to the traditional clinical definition based on the number of human participants.

At the same time, one could argue that the ideal design is not only about capturing more behavioral data, but about ensuring that these measures, learning processes, and robot behaviors remain transparent, interpretable, and clinically meaningful. Automatic outputs may be a powerful aid, but they should be framed in ways that clinicians can trust, translate into established therapeutic language, and ultimately oversee in practice. Transparency, interpretability, and alignment with clinical protocols and operating procedures therefore remain essential goalposts for ASD–HRI research.

Outcome Measures

The field has both diversified and standardized its approaches to outcome measurement. Early studies often relied on qualitative observations (e.g., "the child smiled more with the robot" [233, 278]) or custom metrics lacking baselines or clinical validation. In contrast, more recent work increasingly incorporates standardized instruments and objective behavioral coding. For example, parent- or teacher-report tools like the Social Responsiveness Scale (SRS) are now commonly used to assess changes in social behavior across everyday contexts (e.g., [336–338]). Some studies also report clinical diagnostic scores, such as ADI-R or ADOS, before and after the intervention (e.g., [339–341]), although the sensitivity of these measures to short-term change remains limited [342, 343].

Increasingly, researchers favor proximal behavioral measures that directly capture interaction quality, such as the frequency of eye contact, verbal initiations, or time spent in joint attention. For instance, in Huijnen et al.'s Kaspar mediation study [274], detailed micro-behavioral coding was used to quantify non-verbal imitation, physical contact, attention span, and distraction during robot- versus teacher-led sessions. This fine-grained analysis revealed significantly longer attention spans and lower distraction rates in robot-mediated sessions. Such objective and specific metrics enhance the credibility of claims regarding therapeutic effectiveness although they remain decoupled from clinically-recognized assessment.

Similarly, several studies have sought to quantify engagement through objective behavioral metrics—such as how frequently a child interacts with the robot, how quickly they respond, or how long they maintain attention across sessions. For example, Jain et al. [344] examined engagement trajectories across more than 20 sessions and developed predictive models to identify when a child was beginning to lose interest. The importance of engagement as an outcome measure is increasingly emphasized in SAR literature, given evidence that sustained engagement correlates strongly with learning outcomes.[17] As a result, many interventions now explicitly target the maintenance of social engagement, such as through proxy measures of eye contact, turn-taking, and shared attention, as key therapeutic goals alongside traditional skill acquisition.

However, operationalizing engagement remains a complex and ongoing challenge. Much like basic emotions of anxiety or frustration, engagement can manifest in highly individualized ways and may not always be easily inferred from surface-level behaviors. A nervous child may show a furrowed brow or a forced smile, and a disengaged child may appear still and compliant. For children with ASD, who may avoid eye contact or exhibit atypical gaze patterns, engagement might instead be expressed through body orientation, consistent verbal participation, or physical proximity to the robot. These variations highlight the limitations of relying on surface-level behaviors as proxies for internal states. Consequently, the selection of outcome variables that best reflect true engagement is an active area of discussion in the SAR literature, with no single metric universally accepted as sufficient. This stands in contrast to clinician observation, where the atypicality or nature of an atypical response can often be decoded through a subjective but holistic synthesis of cues.

Crucially, generalization and retention of skills have become the focal outcome of SAR impact, addressing a major gap noted in the 2012 reviews. Earlier, it was unknown if a child who learned to greet a robot would greet a person, or if any gains would persist once the robot was removed. Newer studies are tackling this with follow-ups and human-generalization probes. Scassellati et al. [3] set a high standard by measuring children's joint attention with a clinician outside the robot context. Assessments were conducted a month prior to, during, and a month after a month-long robot intervention. Results showed that children demonstrated improved joint attention with adults following the robot interaction, and caregivers reported fewer prompts and increased spontaneous communication at home. These findings

[17] While engagement is often considered a prerequisite for learning, it does not guarantee skill acquisition. In SAR research, sustained engagement is commonly used as a proxy for attention and motivation, but researchers increasingly acknowledge the need to complement it with direct measures of learning outcomes and generalization.

provided compelling evidence of generalization—the “holy grail” of SAR intervention research—and represented one of the first demonstrations that a robot-taught skill successfully transferred to human interactions.

Retention is another critical outcome increasingly addressed in recent studies, often through follow-up assessments conducted weeks or months after the intervention. For example, a typical trial may measure outcomes immediately after the training phase and then again several months later to evaluate whether children retained the acquired skills or maintained improved social responsiveness. Some SAR studies report encouraging signs of long-term retention. For example, Clabaugh et al. reported long-term retention of intervention content by the children, likely owing to the adaptive nature of the teaching [295]. Similarly, Trombly et al. [345] used an emotionally adaptive robot to teach vocabulary to children with ASD and found that not only did the children learn the target words, but they retained them more effectively than a control group.

In summary, the evaluation of robot-assisted therapy has progressed from predominantly anecdotal and within-session observations to data-rich controlled trial studies. Outcome measures now commonly include: (1) direct behavioral coding of target skills (e.g. instances of joint attention, conversation turns), (2) standardized social ability scales or symptom ratings, (3) measures of engagement (which is both an outcome and a mediator of other outcomes), (4) generalization probes with people and in different environments, and (5) follow-up assessments to check retention. This shift responds to longstanding calls for stronger evidence. As Begum et al. [332] argued, only by following standard clinical research design and showing stable, positive effects can robot-mediated interventions be considered evidence-based practices. The field is moving in that direction, though few studies yet meet the highest standards (e.g., large, multi-site trials). Still, the growing convergence of evidence across recent studies suggesting that robot interventions can improve social functioning marks a notable advancement from a decade ago, when support for such claims was largely anecdotal.

3.6 Discussion

This review summarizes the landscape of robots for autism therapy, tracing the field’s evolution from its origins in 2001 to 2024, consisting of 304 studies. We began by tracing the growth of the field, identifying rapid growth in the number of studies and a shift from exploratory, proof-of-concept prototypes to more structured interventions deployed in real-world environments for long periods of time. We then organized the literature across three connected but discrete phases: designing the intervention goals and structure (Section 3.3); engineering the robot's physical form and behavior to deliver those goals (Section 3.4); and evaluating the outcomes of the robot-assisted intervention (Section 3.5). This framework allowed for a comprehensive examination of the therapeutic, technical, and methodological advancements, as well as persistent gaps and future directions.

3.6.1 A Spectrum of Sociability

Although the positive effects demonstrated by SAR systems are consistently reported in studies that vary in geography, severity of (dis)ability present in users, appearance and capabilities of the robot, and interaction style and setting, there are no clear conclusions on why these robots succeed in establishing and, in many cases, maintaining social engagement.[18] While some behaviors observed during child–robot interactions, such as heightened attention or engagement, can be attributed to the novelty of sensory stimuli, others are less easily explained. Instances of turn-taking with peers [3, 327], spontaneous expressions of empathy [357], or initiative to include the robot in shared activities [358] suggest that robots may occupy a unique space on a spectrum of sociability: more socially evocative than inanimate toys, yet less socially complex than human partners. In this way, robots offer a middle ground: a stimuli for eliciting social behaviors, but not so complex as to overwhelm or confuse children with autism.

There are several compelling hypotheses as to why robots may be uniquely effective for autism therapy. Perhaps it is the predictability and consistency of robotic behavior that provides a more manageable alternative to the often unpredictable and complex nature of human social interaction. For some, robots may offer a socially engaging experience free from the learned anxieties, judgments, or negative associations that can accompany human contact. It is also possible that the exaggerated and explicit nature of robotic social cues (gestures, gaze, vocal prosody) serve as clearer scaffolds for eliciting social behavior than the subtle and often ambiguous signals used by human partners. Perhaps robots offer a form of "social neutrality," wherein their absence of social status or prior relational history makes them less intimidating and more approachable than peers or adults.

[18] Beyond the review presented in this chapter, numerous other reviews have also examined SAR interventions for ASD across diverse contexts. See, for example: [20, 21, 206, 332, 346–356].

3.6.2 Robots as Scientific Instruments

While we position robots as promising tools for intervention, robots also hold significant value as platforms for scientific inquiry. From tracing the development of social interaction across the lifespan, to examining how social norms are transmitted and transformed across generations, to identifying which social cues are most easily perceived and interpreted, robots offer unique experimental testbeds for understanding human social cognition. By embedding computational models of cognition into physically embodied robots and placing them in structured interaction scenarios, researchers can test, evaluate, and iteratively refine these models in real time.

Once embodied, these models can be evaluated in the same types of studies traditionally used to investigate human social behavior, enabling direct comparisons between robot and human responses under similar conditions. Discrepancies between the two provide valuable feedback for improving our understanding of human cognition and inspire new lines of theoretical and empirical research. The granular behavioral control offered by robots allows researchers to systematically manipulate parameters (as low-level as as gaze timing, gesture clarity, or response latency; or as complex as physical form factor and speech) and observe their direct effects on interaction dynamics. Moreover, robots are equipped with sensing hardware that can capture fine-grained behavioral data (e.g., eye gaze, motion trajectories, speech latency, and the timing of social cues) with a level of temporal and spatial precision that far exceeds what is possible through human observation or self-report.

These capabilities make robots uniquely suited for studying the mechanisms underlying social behavior, offering insights into human nature that traditional scientific methods have long sought to understand. In this way, robots contribute both therapeutic opportunity and broader scientific understanding of human cognition.

3.6.3 Autism as a Research Lens

Autism provides a valuable context for studying social behavior because it highlights the often implicit rules that guide human interaction. Individuals with autism may have difficulty with skills that are typically automatic for neurotypical individuals, such as interpreting facial expressions, managing turn-taking in conversation, and adapting behavior to fit social context. The way these challenges manifest in autism exposes the complexity of these behaviors and invites researchers to examine them with greater precision.

This perspective carries important scientific implications. First, the study of autism challenges assumptions about the universality of social cognitive mechanisms, helping to refine theoretical models by showing where they fail to generalize. For instance, when a robot designed to elicit joint attention succeeds with neurotypical children but fails with minimally verbal children with ASD, it suggests a need to revisit our understanding of how joint attention is triggered or maintained. Second, as we have seen in our review corpus, autism research can involve the development of alternative metrics and paradigms to measure social behavior. For example, researchers have proposed using gaze-following latencies or gesture synchrony instead of solely relying on infrequent verbal reports collected by a clinician or caregiver. These more automated, continuous, and fine-grained methodologies can, in turn, be applied more broadly to study social cognition across populations and contexts.

Rather than viewing autism solely as a clinical diagnosis, it can thus be understood as a lens through which the fundamental components of social cognition become more observable and testable. Studying how individuals with ASD perceive, interpret, and respond to social cues offers insights not only into atypical development, but also into the underlying mechanisms of social cognition more broadly.

3.6.4 Designing SARs Across the Autism Lifespan

A central theme throughout the review is the increasing sophistication of SAR systems, not only in technical architecture but also in therapeutic intent. SAR researchers are increasingly attuned to clinical expectations and approaches to ASD therapy. This is reflected in the predominant, almost exclusive focus on childhood interventions. In both clinical and SAR-based therapy, these early interventions feature explicitly modeling, prompting, and reinforcing a targeted social behavior in structured, repeatable ways. It is for these aspects of intervention pedagogy that robots hold such unique promise for effective delivering ASD therapy. Robot-based therapy has addressed a wide range of behavioral targets—from joint attention and imitation to reciprocal language and emotional expression—demonstrating promising therapeutic gains (Section 3.3.2), some of which generalize to human interactions without requiring the robot's presence and persist weeks beyond the intervention period (Section 3.5).

While the reviews published in 2012 underscored the field's near-exclusive focus on children [20, 206], the past decade (2013–2024) has brought a notable expansion in user demographics. Since the initial 55 studies identified prior to 2012, the field has grown by an additional 249 studies. With this expanded evidence base, it is now possible to identify more nuanced trends in SAR therapy for autism. For example, interventions targeting early childhood often emphasize gaze-related skills, such as establishing and maintaining eye contact or joint attention, whereas those focused on middle childhood increasingly address language-based outcomes, including emotional expression and peer communication during collaborative activities.

Despite well-documented evidence that challenges (social, emotional, and functional) often persist or even intensify during later life stages [359–361], relatively few studies have explored how SARs can effectively support adults with ASD. This gap is further surprising given that the first generation diagnosed under broadened diagnostic criteria is now entering midlife, creating an urgent need for developmentally appropriate, scalable supports that extend beyond childhood. Yet, the few existing studies in this space tend to replicate the pedagogical frameworks of childhood interventions, often focusing on foundational skills such as eye contact and emotion recognition, rather than addressing the more complex social or functional needs characteristic of later developmental stages.

There is reason to believe that pedagogical approaches designed for children may not readily translate to meet the needs of adults. Early intervention is based on the idea that there are critical windows in early development—periods when the brain is especially flexible (or "plastic") and able to learn certain skills more easily. Even though scientists do not fully understand the functional importance of these windows (e.g., why they happen exactly when they do), the fact that their timing is so precisely controlled in development suggests they are essential for healthy growth. As the brain matures, one may argue there is a tradeoff between adaptability and stability [195]. The young brain must dynamically adapt to its environment in order to set up its circuits in the most efficient manner while the adult brain instead favors reliability, often resisting change and reacting more conservatively to novel stimuli.

This could mean that adults are less likely to use robots for skill development in the same way children do. They may not benefit from highly repetitive, behaviorist teaching strategies or rigidly structured sessions designed to scaffold foundational social behaviors. Instead, effective adult-oriented interventions may need to incorporate more collaborative, context-sensitive, and internally motivated learning models. For example, rather than teaching isolated social skills, SARs for adults may be more effective when supporting reflective practices, simulating complex social scenarios, or helping individuals navigate specific real-world challenges. Given the limited evidence base for SARs targeting adolescents and adults, the case for their effectiveness in social skills therapy remains far less substantiated than it is for children.

The pedagogical models developed for children emphasize themes of structured learning, behavioral shaping, and scaffolded generalization of discrete skills. These models may fail to translate effectively into adulthood, when social expectations, contexts, and cognitive priorities shift significantly. As individuals with ASD transition into adulthood, the social demands they face become more complex, ambiguous, less easily scripted, and less forgiving of atypical behavior. Adults are expected to navigate dynamic social spaces—such as workplaces, romantic relationships, and public institutions—without the consistent presence of a therapist or caregiver to prompt or reinforce behavior. The cues for appropriate social interaction become subtler, expectations for reciprocal communication rise, and the cost of social missteps may carry more serious social or professional consequences [362, 363].

While recent efforts have begun to acknowledge these gaps, designing SARs that can effectively support individuals with ASD across the lifespan remains an open and underexplored frontier. Future work must extend beyond child-centric paradigms to develop developmentally appropriate interventions that address the shifting cognitive, emotional, and social needs of adolescents, adults, and older individuals. This includes designing robots that can engage users in more abstract reasoning, support identity formation, foster autonomy, and facilitate complex social relationships. Longitudinal research is also needed to understand how interaction patterns and therapeutic needs evolve over time, and how robots can adapt accordingly. Importantly, future systems should emphasize personalization, user agency, and integration into naturalistic routines—principles that may be particularly important for older users who have established habits and preferences. Addressing these challenges presents an opportunity to expand the role of SARs beyond early intervention and toward a lifespan-oriented model of social support.

3.6.5 Moving Beyond Functional Homogeneity

A notable trend across the reviewed literature is the limited diversity in participant functional profiles. The majority of SAR studies focus on individuals described as "high-functioning." The inclusion and exclusion criteria of the studies in our corpus clearly reflect this bias, whether explicitly—by requiring participants to be verbal, have typical vision and hearing, or sustain attention for extended periods—or implicitly through study tasks that demand verbal responses, fine motor control, or sustained engagement. Although this research is nonetheless important, the resulting participant samples do not align with our understanding that is autism is a profoundly heterogeneous spectrum.

The narrow focus on high-functioning participants has implications for both intervention design and broader scientific inquiry. First, it limits our understanding of how specific robot-assisted strategies may need to be adapted for users with different communication or cognitive profiles. Studies rarely stratify results based on participant functional characteristics (e.g., verbal fluency, ASD severity level), making it difficult to evaluate how intervention effectiveness varies across subgroups. Second, this lack of diversity among participants limits the broader scientific utility of robots for investigating the mechanisms of social cognition, as discussed in Section 3.6.2. Interventions involving participants with more complex profiles could help identify which social cues are accessible, which behaviors are challenging to interpret or produce, and how interactions evolve under different cognitive and sensory conditions.

Addressing this gap will require more inclusive recruitment strategies, flexible experimental protocols, and system designs that can accommodate a broader range of user needs. Expanding participation in this way is necessary to ensure that SAR systems are representative of the populations they aim to serve.

3.6.6 What Makes Robots Effective Therapeutic Partners

Earlier in our discussion (Section 3.6.1), we summarized possible reasons as to why robots elicit the positive outcomes consistently reported in prior reviews and throughout the field. First, robots have a physical presence that enables embodied, socially situated interactions in a way that screen-based interventions cannot replicate. Second, robots occupy a middle ground on the spectrum of sociability: more engaging than inanimate toys, yet less complex and demanding than human partners. This hypothesis is not new and has been explicitly examined in controlled studies that compare robot-based interventions to those involving humans [364], animals [365], inanimate objects [366], screen-based avatars [367], or disembodied voices [368–370]. Yet, there remains no consolidated hypothesis explaining why robots elicit the positive outcomes observed in autism therapy.

This review highlights the diverse and nuanced design considerations researcher make in order to develop robot-based therapy for individuals with autism. These span decisions about the robot's physical form, its behavioral repertoire, the therapeutic context in which it must operate, and more. Across all of these dimensions, researchers have adopted highly varied approaches—so much so that no single feature, or combination of features, can yet be identified as the definitive reason *why* robots function effectively as therapeutic partners. To conclude this review, we offer a preliminary hypothesis to guide future investigation into this.

To start, if we abstract the core interaction sequence—from the moment a child notices the robot, forms an impression, initiates engagement, and ultimately chooses to sustain interaction—we can begin to disentangle what specifically drives therapeutic engagement. Each of these interaction phases (from initial attention to sustained engagement) rests on how the user interprets and assigns meaning to the robot's presence and behavior. Understanding this interpretive process is essential, as it underpins how a robot becomes socially relevant and behaviorally meaningful to the user.

We hypothesize that interacting with a robot, even in simple toy-like forms, may involve the rapid formation of a *symbolic framework* through which the robot's role, capabilities, and intentions are interpreted. While not always involving explicit abstract reasoning, this process may rely on a foundational form of symbolic representation that enables users to treat the robot as a socially relevant and behaviorally responsive entity. These initial interpretive steps (though often implicit) may be critical for engagement and interaction and could help explain, in part, the suitability of robots for autism therapy. We organize this hypothesis into four distinct stages: (1) the initial classification of the robot as an object of social relevance; (2) the attribution of agency and intentionality, a process we term *inferential* symbolism; (3) the construction of a model for social participation (*emergent* symbolism); and (4) the development of behavioral expectations based on that model (*predictive* symbolism).

Object Classification. The first act of interpretation may be classifying the robot as something more than an inanimate tool. Despite being mechanical in construction, the robot's motion, interactivity, and often anthropomorphic or zoomorphic features could more easily prompt individuals to distinguish it from conventional objects. This act of categorization could reflect an early symbolic mapping: "robot" is set apart from "furniture" or "toy," and is instead treated as a category with potential for contingent response. For children with ASD, who may struggle with abstract social inference, this visible distinction (what the robot most immediately looks and behaves like) may help facilitate a quicker and more confident classification.

Inferential Symbolism. Even without explicit knowledge, the very act of observing a robot's motion or perceived "response" (e.g., turning its head, making a sound) could trigger an inferential symbolic process. The individual begins to attribute agency (the capacity to act independently) and rudimentary intentionality. This is not necessarily a conscious, explicit thought of "the robot intends X," but rather a pre-conscious, symbolic assignment of “actor” status, distinct from a passive object. This may be akin to how young children begin to attribute agency to animated objects in play, a foundational symbolic leap. Importantly, this process may not require advanced theory of mind skills. Instead, we suggest that the robot’s mechanical consistency and exaggerated responsiveness may make such inferences more accessible for children with ASD, who often benefit from clear, predictable cues in social interactions.

Emergent Symbolism. We now propose that the interaction with a robot inherently prompts the development of a rudimentary symbolic schema for social capacity. While the upper and lower bounds of this capacity for the robot are indeed unknown upon first sight, the very act of engaging with the robot—whether through mimicry, turn-taking, or responding to verbal cues—requires the individual to symbolically represent the robot as possessing some degree of social presence or responsiveness. This symbolic framing does not require an understanding of complex human social norms, but rather a symbolic categorization of the robot as something that can participate in a social exchange, however limited. Even without complex language or nuanced behavior, a robot can be symbolically categorized as socially participatory. For individuals with ASD, who may experience difficulty forming or accessing symbolic representations of “social others” [371,372], mere exposure to robotic interaction could represent therapy in itself. To this point, the robot represents a manageable and consistent “other” for practicing symbolic social exchange.

Predictive Symbolism. We hypothesize that, as interaction unfolds, individuals begin to form symbolic associations between the robot’s actions and their own responses, as well as between the robot’s appearance/features and its potential behaviors. For example, a flashing light might symbolically represent “attention-seeking,” or a specific sound might symbolize “ready for interaction.” This continuous process of predicting and anticipating the robot’s actions relies on the rapid construction and refinement of these predictive symbols. In other words, through repeated exposure users learn to anticipate the robot’s actions with increasing accuracy. Such anticipatory cues may be particularly useful in autism interventions, where many children benefit from highly structured, pattern-based social learning environments.

Taken together, this hypothesis proposes that robots function as valuable partners in autism therapy because they evoke a layered symbolic interpretive process. The gradual progression of symbolic understanding—from simple categorization to predictive modeling—mirrors the progression many therapeutic interventions might aim to support.

At the same time, an opposing argument can be made: the symbolic demands that we hypothesize robots naturally elicit may hinder rather than support engagement. Symbolic play (as we first described in Section 3.3.2) is an already cognitively demanding activity for many children with ASD, as it requires them to understand and accept non-literal uses of objects or actions, track a shared imaginative frame, and interpret the intentions and emotional expressions of their play partner. When the play partner is a robot (that is, a novel agent with novel social cues), one may expect the cognitive load to classify and interpret it would increase. The child must simultaneously decipher what the robot is doing, determine whether its actions are meant to be pretend or literal, and figure out how to respond in a way that aligns with the imagined scenario. This layered complexity can easily overwhelm the child, reducing the likelihood of sustained engagement with the robot-based therapy.

In response to this argument, it may be useful to compare the symbolic demands different agents (e.g., toys, humans, robots) place on users alongside the measured outcomes they produce in autism therapy. In order to do this, future work can revisit the two longstanding motivations for using robots in therapy: their embodied physical presence and their position on the sociability spectrum. For example, what symbolic processes are involved when interacting with a disembodied voice, and how do these relate to observed therapeutic outcomes? How do these processes and outcomes compare when the agent delivering therapy is a dog or an unfamiliar adult? Such comparisons can help empirically establish and clarify the unique advantages that robots offer in autism therapy.

Still, we emphasize that the success of robots for autism therapy does not rely on a single mechanism, but rather on a careful orchestration of interdependent design choices. As described throughout this review, SAR research has treated the full intervention context as a series of careful design choices: What behavior therapy should target and how? Where therapy should take place? Who should be involved in the therapy? How long and how frequent should sessions be? What level of adaptation is needed? What should the robot look like? How should it behave? How should success be evaluated? Although the positive effects demonstrated by SAR systems are consistently reported in studies that vary in geography, severity of (dis)ability present in users, appearance and capabilities of the robot, and interaction style and setting, there are no clear conclusions on why these robots succeed in establishing and, in many cases, maintaining social engagement. The symbolic framing hypothesis offers one plausible account and represents an initial theoretical lens through which future research can investigate this question.

3.7 Summary

In this chapter, we provide a comprehensive overview of the landscape of robot-assisted autism therapy, tracing the field's evolution from its inception in 2001 through 2024 and encompassing 304 studies. Early research in this domain was largely conducted in supervised, controlled laboratory or clinical settings, where one child and one robot would engage in a single session spanning a few minutes. In contrast, the past decade has seen a shift toward more ecologically valid deployments, with robots operating autonomously and without supervision in participants' homes over the course of multiple days or even weeks. However, such studies remain ambitious exceptions, and short-term, dyadic interactions conducted in controlled laboratory settings continue to represent the standard in the field. Moreover, the literature remains predominantly focused on childhood interventions, revealing a significant gap in understanding how robots might support social and functional outcomes in adulthood.

These trends directly inform the core motivations of this dissertation and are reflected across the following chapters. Through five distinct studies, we investigate how robots can be deployed for unsupervised, fully autonomous operation to support sustained, long-term interactions with specialized populations in real-world environments. Of the three studies focused on SARs for individuals with ASD, two (Chapters 6 and 8) represent the first in-home interventions designed specifically for adults. Furthermore, both studies target skill domains that had not previously been addressed in the SAR literature. Chapter 8 presents a level of participant heterogeneity not previously observed in the SAR literature, with nearly half of participants exhibiting co-occurring neurodevelopmental conditions, more severe ASD symptomatology, or limited verbal fluency.

Chapter 4

Challenges Deploying Robots During a Pandemic: An Effort to Fight Social Isolation Among Children

This chapter presents the development and deployment of VectorConnect, a robot telepresence system designed to support remote physical play and social interaction among elementary school-aged children during the COVID-19 pandemic.[1] Motivated by the widespread social isolation experienced during global lockdowns, we sought to create a system that went beyond traditional video conferencing by enabling embodied, peer-to-peer engagement through an accessible robot platform. We review relevant background on the developmental importance of physical play, prior research on robotic telepresence for social engagement, and the unique challenges of designing safe, accessible, and engaging technology for children. We then describe our design goals, implementation, and real-world deployment of VectorConnect, concluding with lessons learned and reflections on how robot-mediated play can foster social connection during times of crisis.

4.1 Introduction

Humans are inherently social beings that rely on interactions with others. It is through these social interactions that we learn, we cope with stress, and we become productive members of society.

The COVID-19 pandemic, declared a global health emergency by the World Health Organization in early 2020, marked an unprecedented period during which nations around the world implemented large-scale lockdowns to curb the spread of a virus. For

[1]This chapter is adapted from our published work: Tsoi, N., Connolly, J., Adéníran, E., Hansen, A., Pineda, K.T., Adamson, T., Thompson, S., **Ramnauth, R.**, Vázquez, M. and Scassellati, B. (2021, March). Challenges Deploying Robots During a Pandemic: An Effort to Fight Social Isolation Among Children. In the *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 234-242). [36]. Additional context and commentary are provided to support its inclusion in this dissertation.

Figure 4.1: Remote Social Play with VectorConnect. A child (right) engages in remote physical play using our system, which enables real-time control of and communication through a Vector robot located in their peer's environment (left). The system facilitates physically playful and socially interactive experiences across geographic distance.

the first time in modern history, billions of individuals were confined to their homes as schools, workplaces, and public venues closed in accordance with emergency health mandates. While these social distancing measures were essential for protecting public health, they conflicted with our instinctual drive for human social connection. Prolonged isolation during this period significantly exacerbated experiences of loneliness [373], a societal issue that had already been identified as a significant and growing public health concern prior to the pandemic [374].

During this lockdown period, many schools and social institutions rapidly transitioned to remote platforms such as video conferencing, online classrooms, and digital messaging platforms. While adolescents and adults possess more developed cognitive and emotional resources for coping with this abrupt transition to digital communication, elementary school-aged children (ages 5-12 years old) are more likely to experience the deleterious effects of social isolation [375]. At this developmental stage, children are still acquiring foundational social and emotional regulation skills and may lack the cognitive maturity, attention span, or motivation to meaningfully engage through electronic platforms. Moreover, face-to-face interaction and physical play are not only central to healthy peer relationships but also serve as critical mechanisms for developing social cognition, cooperation, and self-regulation—essential social competencies that contribute significantly to life-long achievement and social functioning [376, 377].

The COVID-19 pandemic revealed several domains in which robotics could play a critical role, including clinical care, logistics, and surveillance. However, the vast majority of deployed systems during this time were designed to address functional objectives—such as disinfecting environments, delivering supplies, or monitoring public spaces—almost exclusively focusing on physical assistance and operational efficiency. However, despite a well-established body of literature demonstrating the social value of robots, their potential to address the distinct social and emotional consequences of the pandemic remained largely unexplored [224].

To begin addressing this gap, we investigated how robotics might support children experiencing social isolation during the pandemic. Our aim was to explore the capacity of robots not just as functional tools, but as socially interactive agents capable of mitigating some of the developmental and emotional challenges faced by children in prolonged isolation. We developed a robot teleoperation system called VectorConnect to enable elementary school-aged children to engage in physical play despite geographic separation. Built on the widely available commercial Vector robot, VectorConnect allows two users to video chat while one remotely controls a robot located in the other user's environment. Developed as part of our outreach initiative during the first months of the pandemic, our system exemplifies the potential of robots to foster new means for individuals to engage creatively with each other.

We released our system to the general public free of charge. Within three months, approximately 2,000 unique users installed VectorConnect. Although many engaged with specific components of the system's overall functionality, data logs indicate that around a hundred users consistently employed VectorConnect to socialize remotely through the robot. These findings, along with user feedback, underscore the potential of telepresence robots to help mitigate the social impacts of a global pandemic by offering an engaging, safe, and playful medium for remote social interaction.

The remainder of this chapter provides relevant background for the project and presents our experience as a case study in HRI practice. We discuss the challenges that we encountered through our deployment as well as the lessons that we learned to facilitate similar future efforts. It is our hope that this work serves as an inspiration for robotics innovation during times of global crisis.

4.2 Background

The COVID-19 pandemic brought about unprecedented disruptions to daily life, particularly for children whose social, emotional, and cognitive development is deeply tied to in-person interaction. As schools closed and communities implemented social distancing measures, children were abruptly cut off from peers, routines, and opportunities for play—core contexts in which early social skills are formed. These shifts intensified longstanding concerns around childhood loneliness and developmental delay. In parallel, the pandemic also renewed interest in the role of technology—especially robotics—as a means of maintaining social connection under constrained conditions. The following sections overview the effects of social isolation in childhood and examine prior efforts to use robot telepresence systems to support remote social interaction.

4.2.1 Social Isolation in Children

Social isolation presents significant risks to children, both emotionally and developmentally. First, it can lead to loneliness, which has been linked to poorer cognitive functioning and increased mortality risk across the lifespan [378, 379]. Second, it deprives children of the opportunities for interaction and physical play that are crucial for the development of social skills, cooperation, and communication [380, 381]. Physical interaction and peer engagement during childhood are foundational for social-emotional learning, and prolonged isolation can hinder these developmental processes [382–384].

Within the human-robot interaction (HRI) literature, prior work on loneliness has largely focused on how feelings of isolation may influence perceptions of robotic agents. For instance, individuals experiencing higher levels of loneliness have been shown to attribute greater social presence and anthropomorphic qualities to robots [385, 386]. These findings suggest that socially assistive robots may be uniquely well-positioned to engage with users on an emotional and social level.

A particularly relevant study by Odekerken-Schröder et al. [373] analyzed social media posts describing people's interactions with the Vector robot during the COVID-19 pandemic. Their findings suggest that individuals perceived Vector as a comforting presence during periods of social isolation. Similarly, Martelaro et al. [387] demonstrated that expressive robot behaviors can foster trust and a sense of companionship in human users, even in brief interactions.

Our work builds on the broader field of Socially Assistive Robotics (SAR), which focuses on designing robots that support users through social—not physical—means [388]. However, in contrast to systems that serve as stand-ins for human connection, our goal is to use robots to actively connect people, particularly children, who are geographically separated. Specifically, we aim to enable shared play experiences, which are critical to social development and well-being, rather than to offer the robot itself as the source of companionship.

Traditional video conferencing platforms (e.g., Zoom [389]) allow people to connect virtually, but fall short in replicating the physical and spatial dynamics of in-person engagement, which are particularly important for children. Prior work in HRI suggests that physically embodied robots can foster more engaging and positive user experiences than virtual agents alone [390]. These insights motivated our exploration of robot embodiment as a medium for remote physical play among children—a feature lacking in most current telepresence solutions.

4.2.2 Telepresence

Our system leverages a rich history of telepresent HRI. Much of prior work has focused on applications in remote employment, healthcare, and aging, with encouraging results. For example, telepresence robots have been used to support independent living and companionship among older adults [391, 392]. Longitudinal studies show that such systems can increase feelings of social connectedness and promote acceptance among users over time [70, 317, 393, 394]. Notably, Cesta et al. [394] explains that sustained engagement with teleoperated robots depends heavily on their ease-of-use, minimal maintenance, and robustness—factors we prioritized in the development of our system.

More recently, researchers have adapted telepresence systems for use with children in educational and therapeutic contexts. These systems have enabled children to attend classes remotely, practice language skills, and interact with peers despite physical absence [395–397]. Tanaka et al. [396] found that children using a teleoperated robot communicated more effectively than those using traditional video conferencing, likely due to the added physicality and spatial engagement provided by the robot. Other studies highlight that telepresence robots allow children to manipulate remote objects and environments, leading to richer interactions and greater engagement than screen-based alternatives [395].

Our work extends these findings by applying telepresence robotics to the unique context of the COVID-19 pandemic, with a focus on recreational social interaction rather than education or therapy. By enabling remote physical play between children using a commercially available robot, our system explores a novel and underutilized space in HRI: robot-mediated peer-to-peer connection during times of crisis.

4.3 Developing a Robot Telepresence System to Fight Social Isolation

4.3.1 Problem Scope

We identified the problem of social isolation as an important challenge for our society, especially for children, during the early days of the pandemic. As we explored potential solutions, we aimed to develop a system that could help children stay meaningfully connected with peers and family members in a way that was not only accessible, but also engaging. We believed that simply replicating video-based teleconferencing was insufficient; instead, the solution needed to support and encourage physical play. By enabling children to interact through play, the system could make remote communication more enjoyable and developmentally beneficial—resembling the kinds of social experiences children typically gain through in-person interaction.

4.3.2 System Design Goals

Our design goals were shaped by both the social needs of children and the practical realities of deploying a system during a global pandemic. First, we aimed to enable pairs of elementary school-aged children to interact remotely in a way that approximated the experience of co-located physical play. We believed that fostering embodied, playful interaction—rather than relying solely on passive video chat—would make the experience more engaging and developmentally meaningful. Second, we prioritized designing the system with safety and privacy in mind, given the young age of our target users. Ensuring secure communication and minimizing data exposure were essential considerations throughout the design process. Third, we recognized that the system would likely be set up and managed by parents or guardians, many of whom might have limited technical experience. As such, ease of setup and intuitive usability were critical. Finally, we needed the system to be genuinely enjoyable for children—playful, interactive, and capable of holding their attention over multiple sessions—to encourage repeated use and sustained social engagement.

4.3.3 Our Solution

Guided by these design goals, we developed VectorConnect, a mobile application that enables children to engage in remote play with friends or distant family members using an affordable, commercially available robot. Through the app, a child can use a phone or tablet to remotely control a Vector robot located in another child's home—effectively allowing them to "become" the robot in that environment. This setup makes it possible for children to play physically interactive games such as hide-and-seek or to collaboratively build and navigate obstacle courses, all through robotic telepresence. In addition to robot control, the application supports live video and audio communication, allowing children to talk and interact socially in real time. By combining physical play with social engagement, VectorConnect supports a critical developmental need for children in this age group: this type of embodied, peer-directed engagement is strongly linked to foundational social and cognitive development during this formative stage of childhood [398].

Physical Interaction Through the Robot

We selected the Vector robot, developed by Anki, for its suitability in child-centered interaction: its compact size, expressive behaviors, and friendly design make it both safe for children and robust enough to withstand rough physical play [399]. At the time of deployment, Vector was widely available on the consumer market at an accessible price point (approximately $200 USD per unit).

To support diverse forms of physical play, our system enabled one child to remotely control a Vector robot located in another child's home. Using our mobile application, the remote user could access the robot's camera, navigation, and animation functions to engage in shared activities. To protect user privacy, the local child—co-located with the robot—was required to grant explicit permission before these capabilities could be activated remotely.

Social Interaction Through the Mobile Application

To ensure ease of use and familiarity, we designed the interface of our mobile application to resemble common video calling platforms such as Zoom and FaceTime. The application allowed users to see and hear one another via live video and audio on a phone or tablet, while simultaneously enabling control of the Vector robot as previously described. Importantly, video and audio streams were transmitted directly between users—without routing through external servers—to preserve user privacy and minimize latency. This peer-to-peer architecture helped maintain a sense of real-time interaction, making it more natural and responsive for children engaging in remote play.

4.3.4 Implementation Details

We implemented our mobile application to align with the design goals outlined earlier, focusing on simplicity, usability, and privacy. The system architecture consists of a back-end responsible for establishing and managing direct connections between two users and the robot, and a front-end that presents an intuitive interface for interaction. Through the application, users can (a) input the necessary information to connect to and teleoperate a local Vector robot during calls, and (b) initiate video and audio communication with a remote user. The following sections describe the implementation of each software layer in greater detail.

The System's Back-End

The back-end of the VectorConnect system was designed to support secure, real-time peer-to-peer communication between two mobile devices and a Vector robot. This communication relied on two core components: a Traversal Using Relays around NAT (TURN; [400]) server and a Web Real-Time Communication (WebRTC; [401]) connection. The TURN server was responsible for establishing a connection between two users using a shared call ID. This call ID functioned as a private key, allowing only users with matching credentials to join the session, thereby protecting the connection from external intruders. After a successful peer-to-peer connection was established through the TURN server, the WebRTC framework was used to stream phone-based video and audio data, as well as robot commands and video, between devices.

Phone video and audio data were transmitted through WebRTC MediaStreams, which allowed both parties to see and hear each other in real time. Robot commands and robot video were transmitted through a WebRTC RTCDataChannel. This setup enabled the remote user to send teleoperation commands to the robot and view its live camera feed, while preserving low latency and responsiveness across the connection.

To control the Vector robot, we used protocol buffer files originally provided by Anki to implement a Google Remote Procedure Call (gRPC) interface [402, 403]. These protocol buffers offered a well-defined and accessible interface to the robot's core API. However, due to the security features built into the Vector hardware and software stack, establishing a working connection between the mobile device and the robot required substantial reverse engineering. We ported the required functionality from Anki's Python APIs to the Dart programming language, allowing the interface to function on both Android and iOS platforms through the Flutter mobile development framework [404].

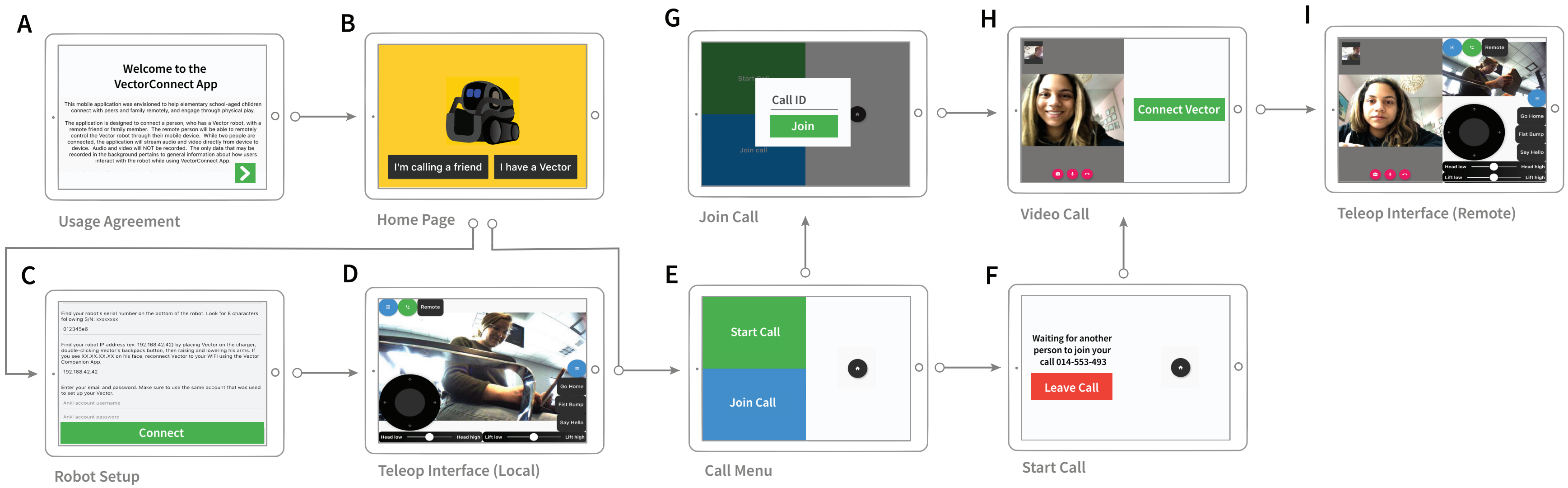


Figure 4.2: VectorConnect Mobile Application Interface. The screenshots above illustrate the user interface flow of our system's mobile application, including (A) the welcome screen with service and data collection policies; (B) home screen for connecting to a robot or joining a call; (C) setup form for secure gRPC connection; (D) local robot control interface; (E) call setup options; (F) generated call ID for new sessions; (G) joining a call using the shared call ID; (H) live video call interface between users; and (I) remote control interface enabled after local user approval.

To obtain a working gRPC connection, the system first queried an Anki API using the robot's serial number in order to retrieve a device-specific Secure Sockets Layer (SSL) certificate. The system then used a valid `anki.com` login, entered by the user, to retrieve a security token from Anki's servers. These credentials enabled the mobile device to establish a secure and authenticated gRPC connection with the local Vector robot. Once the connection was active, the application could control the robot from either a local or remote device. The functionality included sending commands for navigation and animation, as well as receiving sensor data such as the robot's camera feed.

The System's Front-End

The front-end of the VectorConnect application was implemented in Dart [405] using the Flutter mobile development framework. This allowed us to render the interface on both the Android and iOS platforms using a shared codebase. A key goal of the front-end design was to create an interface that was simple and accessible, both for children and the parents who would be helping them set up and use the system. Figure 4.2 illustrates the main user flow.

When the application is first launched, users receive a welcome message with our Terms of Service (Figure 4.2-A). This message explains that the only data collected in the background relates to general app usage and how users interact with the robot. Users must agree to these terms before continuing to the main interface of the application.

The home page of the application (Figure 4.2-B) presents users with two primary options: to set up a connection with a Vector robot or to start a video call with a friend. Selecting the first option brings users to a form interface (Figure 4.2-C), where they are prompted to enter the login credentials for their Anki.com account, as well as the serial number, name, and local IP address of their Vector robot. If the connection is successful, users are taken to a page (Figure 4.2-D) where they can control the robot using a joystick, sliders, and buttons. These controls allow users to move the robot, adjust its head tilt and lift height, and trigger preset animations.

Selecting the video call option from the home page (Figure 2E) allows users to either start a new call (Figure 4.2-F) or join an existing one (Figure 4.2-G). Starting a new call generates a 9-digit call ID, which users can share with a friend to initiate the peer-to-peer session. Once both users are connected and the video call is active (Figure 4.2-H), they can see and hear each other through the video interface of the application.

During a call, the local user—who is physically co-located with the Vector robot—has the option to grant teleoperation access to the remote user. Once permission is granted, the remote user gains access to the robot's camera feed and is presented with the same control interface (Figure 4.2-I) available to the local user. The remote user can then send navigation commands via a joystick, use sliders to adjust the robot's movements, and trigger animations such as the robot motioning a greeting. The interface also includes controls to change the robot's eye color, enhancing its expressiveness during interaction.

Surveys

The VectorConnect application included three optional surveys designed to gather user feedback and demographic information. These surveys targeted both children and parents, aiming to evaluate user satisfaction, understand household demographics, and assess the perceived impact of the system on children's social experiences.

The first survey was a brief, one-question visual satisfaction assessment designed specifically for children. Modeled after the Smileyometer [406], the survey presented five face icons representing varying levels of satisfaction: awful, not very good, okay, really good, and fantastic. This survey was integrated directly into the mobile application and appeared with a 10% probability each time a video call session concluded. Its simplicity and visual format made it accessible to young users and helped us passively collect satisfaction data over time. Figure 4.4 (left) illustrates the child-facing Smileyometer survey.

The remaining two surveys were longer-form and designed for parents or guardians. The first was presented at the initial launch of the application, alongside the Terms of Service. This introductory survey collected basic demographic information for each child in the household, including age, school grade, and gender. It also asked about the child's prior experience with robots—especially familiarity with the Vector robot or other Anki products—and gathered contextual data on the child's school attendance during the COVID-19 pandemic (e.g., whether they were participating in remote learning). In addition, it included questions about the child's recent feelings of loneliness, aiming to assess the social context into which the system was introduced.

The second parent-facing survey was designed to collect ongoing feedback about the household's experience with VectorConnect. This survey included questions about how the child interacted with the robot and the app, what types of play or communication took place, and how parents perceived its impact on their child's mood or behavior. A prompt to complete this survey appeared no more than once per month and was shown after the completion of a video call session, allowing families to reflect on their experience while it was still fresh.

Other Implementation Details

To support robust development and post-deployment monitoring, we integrated a crash reporting system into the VectorConnect mobile application. Specifically, we used Google Firebase Crashlytics, which provided real-time crash reports and deidentified usage data. This integration allowed us to iteratively improve the application by identifying and addressing bugs during both development and real-world use.

As part of our development process, we conducted routine user testing to improve the platform's reliability across a variety of devices. These sessions helped us identify platform-specific crashes and UI inconsistencies. We tested the application across multiple device categories, including iOS and Android operating systems, tablets and smartphones, and hardware released in different years. These tests ensured compatibility and consistent performance across a broad range of devices.

In addition to internal testing, we conducted pilot usability testing with children. In one such session, we observed five children (ages 4–13) from a single household interacting with the system. Each child took turns stepping out of the room with an iPad to remotely control the Vector robot while the remaining children stayed in the room with the robot. During these tests, children first explored the robot's capabilities by driving it and triggering animations. Notably, the in-person children were typically the first to initiate interactive play with the robot, suggesting emergent engagement with its presence.

To guide interaction, we suggested four play scenarios: building an obstacle course, Hide-and-Seek, Simon Says, and Tic-Tac-Toe. During the Hide-and-Seek activity, we described several variants: hiding a physical object for the robot to find, hiding the robot itself, or having a child hide for the robot to seek. In this session, the children experimented with hiding both the object and the robot, but did not engage in the version involving hiding themselves. These observations suggested that the system was engaging and encouraged playful experimentation, even without detailed instructions (depicted in Figure 4.3). Inspired by this experience, we later included these play ideas on the project website as informal suggestions to other users.

To support broader adoption and long-term use, we also created a project website

Figure 4.3: System Usability Testing with Children. In this pilot session, children explored the system's ease of use through suggested and self-directed play scenarios. As shown from left to right, they built obstacle courses using household objects, greeted remote peers via robot fistbumps, played hide-and-seek by hiding the robot, and engaged in a game of tic-tac-toe through remote teleoperation.

(`robotsforgood.yale.edu`) that featured general information, gameplay ideas, and usage tips. In addition, we set up a dedicated support email account to respond to user questions and provide assistance with setup or troubleshooting.

4.3.5 System Deployment

Our initial goal was to release VectorConnect as an official mobile application endorsed and distributed by our university. To this end, we collaborated with the university Information Technology Services to ensure compliance with cybersecurity, privacy, and accessibility standards, including compatibility with screen readers on both iOS and Android platforms. However, due to institutional delays during the early months of the COVID-19 pandemic and the urgent nature of our deployment timeline, we

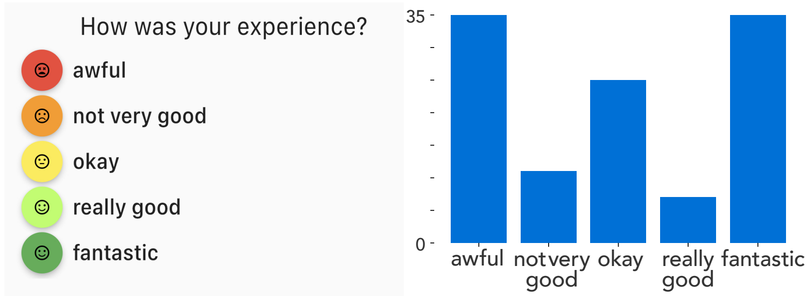


Figure 4.4: Child-Facing Satisfaction Survey Integrated into VectorConnect. The Smileyometer-style survey (left) was shown to users with a 10% probability following each video call. An example response distribution from participants is shown on the right.

were unable to publish the app through official university channels.

Instead, we proceeded with an independent release at the beginning of June 2020, using personal developer accounts from members of our research team to publish the application on the Apple App Store and Google Play Store. The barriers we encountered during this process are discussed in detail in Section 4.5.

In parallel with deployment, we partnered with our university Office of Development to seek donor support for distributing robots to children in need. With generous contributions from the School of Engineering and alumni, we successfully distributed 200 free Vector robots to families in the surrounding community. Section 4.5 also outlines the logistics and challenges involved in acquiring and distributing the robots.

4.4 Results

This project was not motivated to be a traditional HRI study, but rather as an outreach initiative to the ongoing global pandemic. At launch (3 months after the World Health Organization declared the pandemic on March 11, 2020), we promoted the VectorConnect application and project website through a combination of news articles, institutional communications, and social media. As part of this effort, we distributed 200 Vector robots free of charge to families in our local community on a first-come, first-served basis. These distributions were unconditional—recipients were not required to use the application, provide feedback, or participate in any formal study. In line with our commitment to accessibility and user privacy, we intentionally minimized data collection to reduce barriers to participation and maximize community reach.

Importantly, the application was not restricted to families who received a free robot. VectorConnect was made freely available for download to the general public in the United States through both the Apple App Store and the Google Play Store.

The following sections describe user adoption and system performance between June and September 2020. We report anonymized usage data, application ratings submitted both in-app and through public app stores, and crash logs. While no participants completed the optional monthly feedback survey by the end of this period, we were able to collect limited demographic data through the initial onboarding process. Usability insights are discussed in the context of this self-reported information.

4.4.1 User Demographics

As described in Section 4.3.4, demographic information was collected via an optional parent-facing survey displayed upon the application's first launch. The goal of this survey was to gather basic information about the children who would be using VectorConnect.

Between the release of the application in June 2020 and the end of September 2020, 48 parents initiated the demographic survey, and 30 completed it. These 30 completed responses represented a total of 47 children, of whom 27 were male and 20 were female. Forty-one of the children (87%) fell within our target age range of 5 to 12 years old, confirming that the system largely reached its intended user group. One child's age was not reported. The median age was 9, and the mean age was 8.5 years. Among the seven grade levels reported, sixth grade was the most common, representing 10 of the 47 children.

Most of the children were affected by school closures during this time: 77% ($N = 36$) were not attending school in person due to the pandemic. In terms of psychosocial well-being, 85% ($N = 40$) were described by their parents as experiencing some degree of loneliness while at home. Moreover, 94% ($N = 44$) were reported to desire more social interaction with distant peers or family members. Despite this desire, 60% ($N = 28$) were reported to interact with other remote children only once a week or less. One parent noted the lack of meaningful play opportunities, stating: "no real play, [just] talking and texting." Our system aimed to address this gap by enabling richer, play-based remote interactions.

4.4.2 System Adoption and Usage Patterns

In this section, we discuss how users took advantage of our application based on system logs and feedback through the app stores.

Effective Users

There were a total of 1,985 unique users that launched the application from the release in June to the end of September 2020. From this set, 92% (N = 1,828) of unique users accepted the Terms of Service and continued into the application.

Among these users, 91 engaged with both core functionalities of the application: connecting to a Vector robot and initiating a video call with a friend. On average, these users were connected to a Vector robot for a total of 20.37 minutes ($\sigma = 28.18$) and spent 16.18 minutes on video calls ($\sigma = 38.42$). This reporting includes only calls that lasted one minute or longer.

The number of unique users significantly exceeded the 200 robots we distributed to families, suggesting that our system reached a broader audience than our original outreach. As discussed in the following sections, some users adopted the application for alternative use cases—such as remotely controlling their own robot——highlighting the flexibility and unanticipated appeal of the platform.

Connect to Vector

Among the 1,828 users who accepted the Terms of Service, connections were made to 759 unique Vector robots. In total, the system recorded 3,788 individual connection sessions, of which 1,989 lasted at least one minute. The average duration of these connections was 3.56 minutes ($\sigma = 4.31$).

Interestingly, 87% of users (N = 592 of 683 total) only used the application to control the robot and did not use the video calling feature. The average of these users controlled the robot for a total of 8.82 minutes ($\sigma = 16.16$) since the release of the application. These users made an average of 2.9 connections to the robot ($\sigma = 3.7$).

Comments left by users in the app stores also reflected the extent of our application's impact beyond the 200 families that received robots. This impact may have been enhanced by the fact that the remote-control features in our application were unique. For example, one Android user started his review for VectorConnect saying, *"I really recommend this app because it lets you do things that you can't do on the official app, such as controlling Vector and seeing his perspective."* The official "Vector Robot" application by the robot manufacturer did not include any such remote control features. Thus, it is possible that some users may have downloaded our application primarily to teleoperate Vector.

Call a Friend

A total of 6,440 calls were made between June and the end of September 2020. Of these calls, 336 instances lasted one minute or longer. These calls were made by 193 unique users and had an average duration of 8.86 minutes ($\sigma = 35.67$).

There were 102 unique users that used the application only to make video calls, never connecting to a Vector robot. An average user in this category made 1.7 calls ($\sigma = 1.50$). The total duration of all calls for this group averaged 14.75 minutes ($\sigma = 53.99$).

While many users either made video calls without controlling a robot or used the robot without initiating a call, our analysis showed that users who engaged in both activities—controlling a Vector robot while simultaneously video chatting—tended to remain active for significantly longer sessions. On average, these users exhibited nearly double the engagement time compared to those who only used the robot control feature. This suggests that the combined experience of interacting with a friend through video while also collaboratively operating the robot was more immersive and compelling than either feature alone.

User Engagement

To assess how meaningfully users engaged with our system, we analyzed standard user engagement metrics: Daily Active Users (DAU), Weekly Active Users (WAU), and Monthly Active Users (MAU) during the period of late August through September 2020. We calculated the ratios between these metrics to approximate user retention.

Our application had a DAU/MAU ratio of 6.6%, a DAU/WAU ratio of 21.9%, and a WAU/MAU ratio of 30.0%. A DAU/MAU ratio of 6.6% implies that, on average, users engaged with the app on 2 out of every 30 days. As shown in Figure 4.5, these metrics remained relatively stable throughout the measurement period. While benchmarks for retention can vary widely across applications, a study of Facebook apps found median DAU/MAU values between 5.6% and 9.0% for lifestyle, entertainment, and game apps [407], suggesting that our application's retention fell within an expected range for socially oriented consumer apps.

Anecdotally, one of the authors used VectorConnect to play remotely with two children in her family (ages 5 and 8). Although they were prompted to drive the robot through a toy obstacle course, the children quickly discovered more playful uses, such as dismantling the course and changing Vector's eye color. This highlights the emergent, exploratory nature of play that our system was able to support.

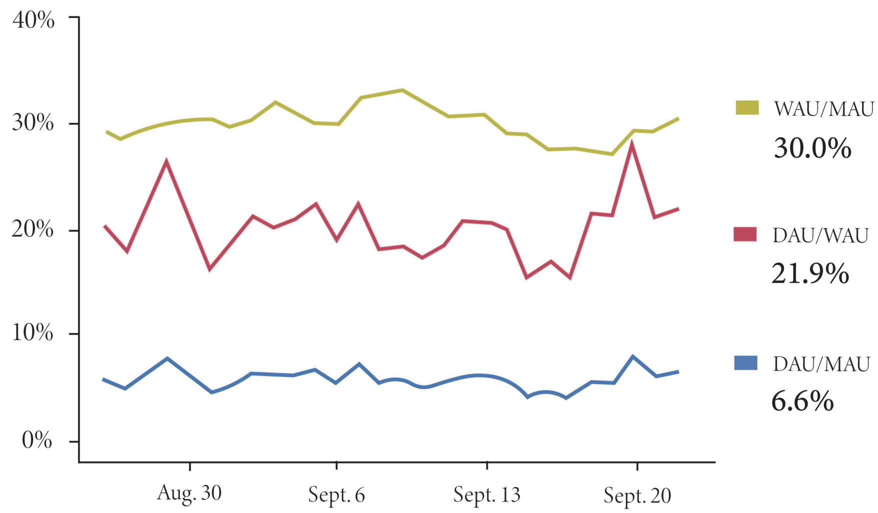


Figure 4.5: User Retention During Data Collection Period. A DAU/MAU ratio of 6.6% indicates that, on average, a user engaged with the application on 2 days per month. Notably, these engagement levels remained consistent from late August to September 2020, suggesting sustained interest beyond initial novelty and across the child-focused user base.

4.4.3 User Satisfaction

The optional smiley face survey offered users a simple, child-friendly way to rate their experience with the application on a 5-point scale ranging from "awful" to "fantastic."" Although many users exited the app immediately after a call without completing the survey, we received 113 responses between the public release and September 2020. The detailed results are shown in Figure 4.4.

The smiley face survey revealed that the majority of users were satisfied, ranking the application "okay" or above ($N = 67$ of 113 respondents), yet responses were polarized. Respondents often picked the extreme values of "awful" or "fantastic," both of which received 35 responses.

Additional insight came from user reviews in the app stores. In September 2020, the Apple App Store rating averaged 3.7 out of 5 (18 reviews), while the Google Play Store averaged 4.0 out of 5 (7 reviews). Users who left positive feedback praised the unique features of VectorConnect, especially the remote control functionality not available in other applications.

Negative reviews primarily referenced technical issues, such as app crashes or difficulties connecting to the robot. Our backend logs revealed that a significant proportion of crashes stemmed from a bug in a third-party library used during development. Despite these issues, the app remained stable for the majority of users, with 90.24% of sessions in September 2020 being crash-free.

Connecting to Vector also proved to be a barrier for some users. As described in Section 4.3.4, the setup process required entering detailed information—such as the robot's IP address, serial number, and Anki login credentials—which was often difficult for children to provide independently and sometimes unclear or error-prone for parents. Although we provided a step-by-step online guide, simplifying this setup flow could further reduce friction and increase adoption.

Finally, we observed video latency issues on older mobile devices when streaming both the robot's video and the friend's video simultaneously. While our website listed recommended device specifications based on internal pilot testing, we did not restrict downloads from unsupported devices. This openness helped maximize accessibility but may have contributed to performance inconsistencies, particularly for users with outdated hardware.

4.4.4 Summary of Findings

Given the outreach-oriented nature of our project, we did not have a direct mechanism for confirming whether our system measurably reduced feelings of loneliness among users. However, the demographics data presented earlier indicated a clear need: many children were isolated at home and eager for remote social interaction. Usage logs and user feedback suggest that, despite its imperfections, our system was meaningful to a broad user base. Hundreds of video calls were successfully established between remote users, many of which included robot-mediated play. Interestingly, beyond our intended use case, a large number of users adopted the app as a tool to teleoperate their own local Vector, highlighting the broader utility and appeal of our platform.

4.5 Barriers and Challenges

The following sections outline key challenges we encountered from the inception of the project through its deployment. By sharing these experiences, we aim to help future teams anticipate similar obstacles and navigate them more effectively.

4.5.1 A Global Pandemic

The COVID-19 pandemic introduced widespread uncertainty and logistical complexity, and our efforts were no exception [408, 409]. Routine processes that would typically be resolved through a single in-person meeting instead required multiple virtual interactions, often delayed or disrupted.

Our initial plan was to distribute the donated robots through local public schools, which we believed would be an effective channel to reach children in need. However, ongoing school closures and the shifting priorities of school administrators—particularly during the June end-of-year transition—made it difficult to coordinate and secure approval for this approach. As a result, we pivoted to an alternative strategy: distributing the robots directly to families by advertising the opportunity online, through media coverage, and via word of mouth.

4.5.2 Choice of Robot Platform

We selected the Vector robot for several compelling reasons: (1) it met our key design specifications, including safety, expressiveness, robustness, and affordability, as outlined in Section 4.3.2; (2) prior research suggested that Vector has the potential to alleviate feelings of loneliness [373]; (3) it was commercially available at the time of development; and (4) our team had prior experience working with a similar robot, Cozmo. However, the selection was not without complications. Anki, the company that originally developed Vector in 2018, had gone out of business, and the rights to the robot had since been acquired by Digital Dream Labs (DDL). This raised immediate concerns about the longevity of support for the platform and the available inventory, as new units were no longer being manufactured.

Despite these uncertainties, former Anki engineers generously offered technical insights that helped us navigate the platform's limitations, and representatives from DDL confirmed that ongoing support would be provided.

Two major technical challenges emerged: establishing a reliable connection between the robot and our mobile application, and accessing certain internal components of the robot. These difficulties stemmed from Vector being designed primarily as a consumer product rather than a development platform. Consequently, many lower-level features were either undocumented or restricted, forcing us to reverse-engineer parts of the software stack. For example, we had initially hoped to enable remote users to access audio from the robot's microphone, but the relevant functionality was inaccessible through the provided APIs.

While Vector offered a rich set of expressive behaviors, its closed architecture ultimately limited what we could implement. Our experience highlights the need for more socially engaging robots on the market that combine robust design with developer-friendly programming interfaces—especially for researchers and educators looking to create interactive systems for children.

4.5.3 Price Gouging and Seller Approval

During the early months of the pandemic, price gouging was widespread [410], and Vector robots were not exempt from this. This issue was further compounded by the fact that Anki was no longer manufacturing new units, making available robots increasingly scarce. As we explored options for purchasing robots in bulk for distribution, we encountered delays due to the need for institutional approval processes, which were necessary to authorize large expenditures. Unfortunately, the price of the robots rose significantly during this time. In one case, the price increased by 75% between our initial conversations with donors and the final approval to purchase.

By the time a vendor was officially approved, inventory had diminished and market demand remained high. As a result, although donors had committed funds based on initial cost estimates from earlier in the pandemic, the total number of robots we were able to procure ended up being slightly fewer than originally planned.

4.5.4 User Privacy

We implemented a number of security and privacy measures in our mobile application to protect child users. Some of these measures were planned from the outset, while others emerged as necessary during the process of preparing the application for public release. For instance, from the beginning of the project, we decided that each video call would be initiated with a newly generated nine-digit call ID. This call ID was not automatically shared by the app, and instead had to be communicated externally (e.g., by a parent) to the intended recipient. This design choice added a layer of protection by requiring parental coordination before a child could join a call, thereby preventing unsolicited or unmonitored interactions.

To comply with Apple's requirements for children's apps, we also implemented a "parental gate" within the iOS version of the app. This gate took the form of a simple arithmetic problem that users had to solve before being allowed to open external links, such as those leading to our project website or online surveys. Although this feature was not part of our initial plans, it proved essential for getting VectorConnect approved for distribution via the Apple App Store.

Additionally, we considered how to signal when a Vector robot was being remotely controlled. At the time of development, Vector's API did not support changing the robot's backlights, and we chose not to use its screen for this purpose in order to preserve the face customization options available to users. As a result, the robot had no built-in visual indicator of remote control. In future iterations, we believe adding a clear visual signal to indicate teleoperation would enhance transparency and help ensure that users—especially children—remain aware of when the robot is actively controlled by someone else.

4.5.5 Institutional Review Board Approval

Due to the pandemic, our local Institutional Review Board (IRB) implemented new subcommittees to fast-track the review and activation of COVID-related projects. These subcommittees were intended to provide end-to-end oversight, from initial study concept through implementation.

However, the newly established pandemic-specific approval processes were not fully integrated with the IRB's existing workflows, leading to delays and procedural ambiguity. Additionally, institutional resources were heavily focused on urgent priorities such as COVID-19 testing and contact tracing, which further limited the availability of reviewers and substantially slowed our efforts. Our study ultimately received full IRB approval.

4.5.6 Institutional Friction

Our team anticipated going through the processes necessary for creating an application affiliated with our university. Thus, we worked for a significant time to comply with institutional requirements. This included working with the university to ensure that our application met the identity guidelines of our institution and tailoring our development processes to accommodate university requirements to publish our application, such as fulfilling Web Content Accessibility Guidelines. We also proactively mitigated cybersecurity risks to our institution and to our users, developed a privacy policy that comports with our institutional approach to privacy, and demonstrated that our existing privacy safeguards complied with the university guidelines.

However, in exerting its brand control, the University Printer's office had to review our application's icon to ensure that the icon met the identity guidelines of our institution. The process to approve the icon took over three weeks, compounding the delay from the institution's developer team to approve our application. Furthermore, the review by the Office of General Counsel was significantly delayed due to an increased volume of review requests within the Office.

Therefore, our team finally opted to publish the application using private developer accounts instead of our institutional account. Had we decided to publish privately earlier, we would have saved several weeks of delay and a significant amount of effort.

4.6 Opportunities and Recommendations

We hope to raise broader awareness of the potential roles that robots can play in mitigating the social impacts of infectious disease outbreaks. In addition, we aim to support future HRI efforts by sharing key insights from our experience deploying a time-sensitive application. The following sections outline the lessons learned throughout this process.

4.6.1 Procedural Changes

There is a need for coordinated institutional efforts to streamline and strengthen administrative procedures against major disruptions, such as an infectious disease outbreak or other events of crisis. While hardening institutional processes is beyond the direct scope of the HRI community, our project illustrates how procedural hurdles can significantly affect HRI research and deployment. By raising awareness of these challenges, we aim to inspire institutional reforms and collaborative solutions.

In our case, the university introduced an additional layer of approval processes to address pandemic-related concerns. However, these procedures were not well integrated with existing institutional review frameworks. Moreover, our application, despite its relevance, competed with other COVID-related efforts for limited reviewer time and attention. Rather than routing all new initiatives through newly formed channels, institutions could define clear exemption criteria for time-sensitive academic projects. Such foresight would enable innovative technological responses to proceed more efficiently in times of crisis.

4.6.2 Market Opportunities

Our project underscores a clear market opportunity for low-cost, reliable robotic platforms that offer accessible development tools and robust documentation to support third-party innovation. Despite growing interest in HRI, the current consumer robotics market remains fragile. Many promising start-ups in this space operate with limited resources, and as a result, are often short-lived. This volatility makes it difficult to depend on any one platform for sustained deployment, particularly during emergencies when reliability and long-term support are critical.

The challenges we encountered (e.g., limited access to proprietary APIs, uncertainty about long-term platform support, and the need to reverse-engineer components) highlight a broader structural issue. When companies dissolve or are acquired, essential resources like documentation, firmware, and development environments often disappear or become inaccessible. While this may be partially addressed as the robotics market matures, it presents a critical risk in the current state of the field.

We hope that our experience serves as a call to action for robotics companies to preserve documentation and development interfaces in the public domain whenever feasible. Doing so would ensure that academic and humanitarian efforts can continue to build upon these systems, even if the original company ceases operations. The inclusion of open APIs, modular firmware, and clear licensing for continued community use would enable greater resilience and long-term value.

Although we could not have foreseen the pandemic, our experience suggests that earlier partnerships with existing manufacturers—regardless of whether their platforms were a perfect fit—might have improved the stability and scalability of our system. Future initiatives might benefit from establishing such partnerships in advance, especially if the goal is to enable rapid deployment in response to societal needs.

4.6.3 Readiness Initiative

The challenges we faced in deploying our teleoperation system could have been significantly mitigated with the support of a dedicated partner organization capable of collaborating on publicity, outreach, and distribution. One model for such a partnership is the Center for Robot-Assisted Search and Rescue (CRASAR) [411], which advocates for the use of unmanned systems in emergency response and public safety. CRASAR has played a central role in coordinating rapid responses to natural disasters, such as wildfires, floods, and hurricanes, by mobilizing volunteers, disseminating solutions, and acting as a trusted hub for the search-and-rescue community.

A similar nonprofit or consortium focused on socially assistive robotics could greatly amplify the impact of time-sensitive HRI efforts like ours. By providing established infrastructure for technology deployment, such an organization could accelerate access to target populations through broader publicity and more streamlined distribution channels. Moreover, it could help relevant stakeholders, such as schools, public health officials, and families, quickly understand the potential of robotics in addressing urgent societal challenges, including the secondary effects of infectious disease outbreaks [224].

The presence of such a coordinating body would empower research teams to focus on innovation and rapid response, while helping ensure that valuable interventions reach those in need more efficiently. Ultimately, institutionalizing this kind of support could pave the way for more agile, impactful applications of socially assistive technologies in future crises.

4.7 Summary

Social isolation can have profound effects on child development, contributing to loneliness, poorer health outcomes [378], and increased mortality risk [379]. In response to the acute social challenges posed by the COVID-19 pandemic, we developed and deployed VectorConnect—a robot teleoperation system designed to foster meaningful, physically-embodied interactions between children separated by distance. Our system enabled remote users to communicate through video calls while jointly engaging in physical play via a Vector robot.

Despite a limited deployment window, VectorConnect reached hundreds of users across the United States and demonstrated promising potential to support child well-being during periods of extended isolation. While our findings affirmed a real demand for platforms that blend social connection and tangible interaction, they also revealed the considerable barriers that can hinder rapid HRI deployment. These included institutional bottlenecks, market instability in robotics hardware, and a lack of infrastructure for emergency-response HRI initiatives.

Our experience underscores the need for more resilient institutional procedures, accessible and developer-friendly robot platforms, and coordinated community efforts to mobilize socially assistive technologies in times of crisis. We advocate for better preservation of robotics intellectual property and the establishment of dedicated organizations to support the deployment of HRI systems in response to public health and humanitarian emergencies. By sharing our lessons learned, we hope to encourage others in the HRI community to design with urgency, deploy with care, and build systems that can truly make a difference when they are needed most.

As a chapter in this dissertation, this work summarizes the ongoing challenges we faced while deploying socially assistive robots during the COVID-19 pandemic. These challenges—ranging from technical limitations to institutional and logistical hurdles—were not isolated; they re-emerged across three of our other robot deployments that overlapped with pandemic-related protocols. This study shaped our understanding of what is required for effective design, development, and real-world deployment during a time of global crisis. As such, the lessons learned here directly inform our approaches and adaptations presented in our subsequent deployments. In this next chapter, we transition from a robot aimed at fostering general social engagement in children to a robot intentionally designed to support targeted aspects of childhood social development.

Chapter 5

Gaze Behavior During a Long-Term, In-Home, Social Robot Intervention for Children with ASD

The previous chapter explored how robot-mediated play can support broad social needs during a time of crisis, highlighting both the challenges of deploying socially interactive robots in real-world settings and their potential to meaningfully shape children's social experiences. Although that study addressed the unique constraints of in-home deployment during the global pandemic, the in-home intervention analyzed in this chapter was conducted much earlier. Any thematic overlap reflect longstanding questions within our broader research agenda, rather than insights derived from the COVID-era deployment. Here, we turn to a more focused and enduring question that has motivated our work from the outset: how might socially interactive robots support targeted developmental outcomes when used consistently over time in everyday environments?

One population for whom this question is particularly urgent is children with Autism Spectrum Disorder (ASD) as they face persistent challenges in social communication and interaction. This chapter examines the impact of a month-long, in-home, robot-assisted intervention aimed at improving gaze behavior in children with ASD.[1] Appropriate gaze behavior is a foundational component of early social development, a prerequisite for more complex social skills, and a core diagnostic feature of ASD. The intervention, conducted by Scassellati et al., in 2018 [3], was a landmark study that demonstrated both the feasibility and promise of robot-assisted interventions for ASD.

[1]This chapter is adapted from our published work: **Ramnauth, R.**, Shic, F., & Scassellati, B. (2025). Gaze Behavior During a Long-Term, In-Home, Social Robot Intervention for Children with ASD. In the *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 949–957). IEEE. [35]. The original study, including its design, development, and data collection, was conducted by Scassellati et al. and published in 2018 [3]. That study introduced the robot intervention that serves as the basis for the current work. I was not involved in the execution of the 2018 study. Rather, in this chapter, I present a new analysis and interpretation of the 2018 data, conducted independently to explore a different set of questions.

Not only did it validate that such in-home systems could be deployed successfully, it also provided evidence of meaningful developmental gains—most notably, improvements in joint attention. At the time, however, the gold standard for evaluating these outcomes relied on clinician-administered assessments conducted in the home once at the start and once more at the end of the intervention. While this approach yielded promising outcomes, it left several critical questions unanswered: When during the intervention did these behavioral changes emerge? Were they gradual or abrupt? Consistent across participants or highly individualized?

Understanding the timing of behavioral change has important implications for the future of autonomous therapeutic systems. If we can identify *when* behavioral improvements occur, it may be possible to develop systems capable of autonomously detecting those inflection points—recognizing, in real time, when they are effectively supporting users. To reach that goal, we needed to revisit the computational methods for automatically extracting and interpreting behavioral change. In this chapter, we address each of these open questions: Was the SAR-based therapy effective? Did it lead to measurable behavioral improvements? Can behavioral change be automatically and accurately detected from interaction data? When, precisely, did these changes emerge? And, more broadly, what do these patterns reveal about ASD and the design of robot-based interventions for such a uniquely heterogeneous population?

Although thematically related to Chapter 4, the intervention by Scassellati et al. was conducted earlier (in 2018) and our analysis presented here reflects an independent line of inquiry. Rather than focusing on general social engagement during a time of crisis, this chapter centers on targeted, clinically meaningful skill development. In doing so, it offers a deeper understanding of the behavioral dynamics and learning trajectories of children with ASD, highlighting the potential of socially assistive robots to support long-term developmental goals in real-world home environments.

5.1 Introduction

Social interactions involve complex exchanges of gaze. People rely on eye contact to direct attention to objects or events, respond to others' shift in attention [412], encourage prosocial behaviors [413, 414], and infer others' thoughts, desires, or intentions [415, 416]. Recent findings emphasize the key role our gaze patterns play in coordinating joint activities [417] and facilitating social learning [418–420]. In essence, gaze serves a critical communicative function and its temporal dynamics provide valuable cues in a social exchange.

Yet, eye contact in Autism Spectrum Disorders (ASD) is a subject of continuing discussion in the literature. Atypical gaze behavior is a diagnostic hallmark of ASD and contributes to many of the social and communicative challenges individuals with ASD face [242, 421]. For example, individuals with ASD show a reduced motivation to share attention with others [422]. Compared to neurotypicals, individuals with ASD initiate joint attention to a lesser extent, are less sensitive to social gaze, and tend to avoid eye contact [423, 424].

It is commonly believed that training appropriate gaze behavior will enhance one's overall social skills because it is considered a prerequisite for more complex behaviors [425]. Therefore, eye contact is often targeted first for ASD intervention [242]. The intervention pedagogies are typically centered around positively reinforcing naturally-occurring incidences of eye contact [426, 427], modeling eye contact with others during social interactions [428], or adjusting one's behavior by using visual supports to encourage eye contact with a speaker [429, 430]. Although these interventions are intuitive methods for training appropriate gaze behavior, they demand the continued motivation of the caregiver, consistency in their behavioral feedback, and constant sensitivity to the specific needs and abilities of the individual with ASD over time.

Socially assistive robotics (SARs) has the potential to augment the current efforts of caregivers and clinicians by eliciting positive and productive outcomes in ASD interventions [20]. The robots envisioned by these efforts support social and cognitive growth by improving access to on-demand, personalized, socially-situated, and physically co-present interventions. Research on SARs for ASD show increased engagement, improved attention regulation, and more appropriate social behavior such as joint attention and spontaneous imitation when robots are part of the interaction [20, 21].

However, many of these studies focus on short-term interactions under controlled settings, or ultimately fail to demonstrate learning that generalizes to human-directed actions. In response to this critical gap in the literature, Scassellati et al. [3] reports directly assessed improvements in social skills in children with ASD following an in-home, month-long intervention conducted by an autonomous, socially assistive robot. The study is a preliminary step to demonstrating that SARs are capable of producing lasting enhancements in social and communicative skills that are generalizable beyond the specific robot encounter to real-world, human-human interactions.

The rich dataset that resulted from this study characterized skills improvement using standard assessments at four time points: (i) 30 days before the intervention began; (ii) on the first day of the robot intervention; (iii) on the last day of the intervention; and (iv) 30 days after the end of the intervention. These assessments include measures of engagement based on the child's performance in various social skill games, joint attention between the child and their caregiver, and caregivers' surveys of their child's initiation of eye contact and verbal communication beyond the robot-assisted intervention.

Automated measures of performance in dynamic, unconstrained environments like the home demand complex sensing. In the original study by Scassellati et al., gaze behavior was manually assessed by a clinician at these four discrete time points. These assessments were designed to evaluate skill transfer: specifically, whether gaze behaviors learned in the robot-parent-child triad generalized to a child-clinician interaction without the robot present. While these evaluations speak directly to the clinical impact of the intervention, they do not provide insight into how gaze behaviors evolved throughout the course of the month-long deployment. A more continuous analysis may better capture the subtle and nuanced patterns of improvements that unfolded over time. To enable this, a reliable method of automatic gaze extraction must first be developed and then applied to the entire source dataset. The results produced by this automated method will not only confirm the manually coded outcomes but also provide valuable insights into the sensing required to accurately detect gaze behavior in the home.

Furthermore, the original study reported only this single aspect of gaze behavior: joint attention between the child and the clinician. The source data contained substantial information about other forms of gaze behavior such as attentional shifts, mutual gaze, and gaze-following among the robot, child, caregiver, and other agents in the home environment. This additional exploration can provide a more comprehensive analysis of the effects of the intervention on gaze behavior in ASD.

In summary, Scassellati et al. [3] presented initial findings on the feasibility and efficacy of delivering an ASD intervention with a robot. This study expands the definitions, detection, and analysis of gaze behavior in [3] to better describe the effects of a long-term, in-home, social robot intervention on gaze behavior in ASD. The results of this expanded analysis have the potential to transform how we design and approach long-term, robot-assisted interventions for ASD.

5.2 Method

Participants for the initial study were recruited through the university's medical school, and the current research team obtained Institutional Review Board approval to access their data. The following sections provide details on the participants, the design and components of the intervention system, and the methods used to extract and analyze gaze behavior from the interactions. Most of these details were not included in the initial study by Scassellati et al., thus supplementing the prior work and providing essential context for the current analysis.

5.2.1 Participant Information

Fifteen families with a child with ASD enrolled in the study. Two families withdrew, one due to unrelated health problems and one due to technical difficulties with the robot installation. Among the families who completed the study, five of the children were females and eight were males. The participants' age ranged from 6 to 12 years old ($M = 10.0$, $SD = 1.4$).

The diagnosis of ASD was established using a clinical best-estimate (CBE) approach by licensed psychologists and speech-language pathologists experienced in the diagnostic process. Prior to study inclusion, autism symptoms were characterized using the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS), gold-standard tools for clinical ASD diagnosis. The ADI-R is a semi-structured parent interview assessing four domains: reciprocal social interactions ($M = 18.3$, $SD = 6.9$, cutoff: 10), communication ($M = 16.6$, $SD = 4.7$, cutoff: 8), restricted and repetitive behaviors (RRB; $M = 3.6$, $SD = 0.8$, cutoff: 3), and early abnormal development history ($M = 3.4$, $SD = 0.7$, cutoff: 1). The ADOS is a clinician-administered behavioral assessment that provides a calibrated severity score ($M = 7.3$, $SD = 2.0$, cutoff: 4). While useful for symptom characterization, both tools were designed for diagnostic classification, not for measuring change, as they lack item granularity and sensitivity/specificity to short-term intervention outcomes [431,432]. In this study, the ADI-R and ADOS were used to explore associations between detected changes and ASD symptomatology.

All participants had IQ scores ≥ 70 as measured by the Differential Ability Scales (DAS-II), with means across verbal reasoning ($M = 91.8$, $SD = 25.9$), nonverbal reasoning ($M = 95.2$, $SD = 15.7$), spatial reasoning ($M = 94.2$, $SD = 16.0$), and general conceptual ability ($M = 93.1$, $SD = 19.6$). Participants exceeded ASD cutoffs

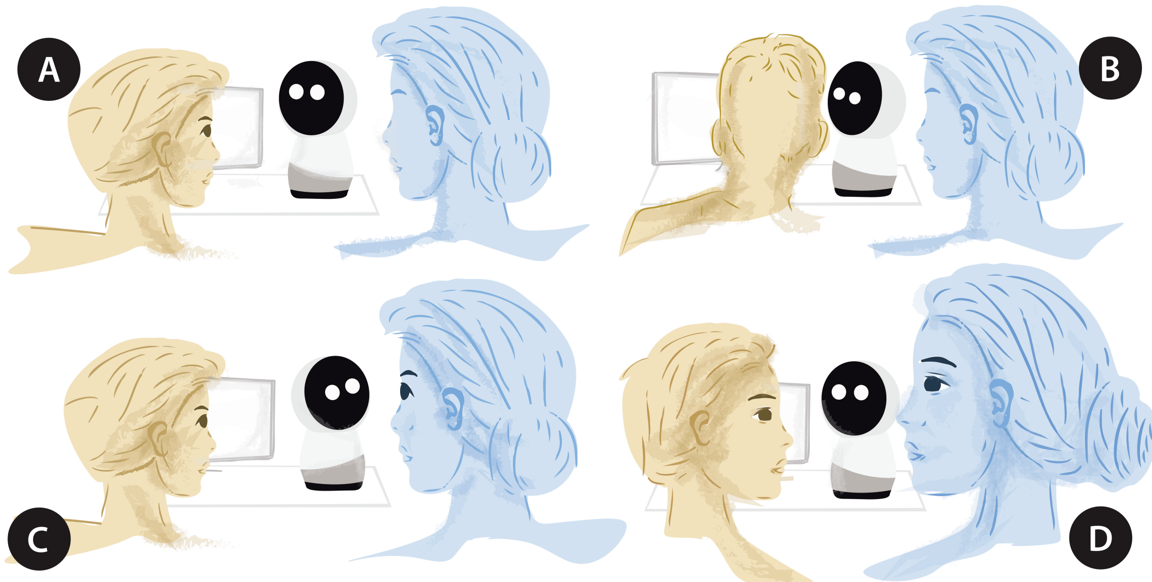


Figure 5.1: Modeling Gaze. The robot's context-contingent gaze guides the child's attention between the screen and caregiver, promoting increased interaction. When the child looks at the robot (A), it first redirects the child's attention to the game content on the screen (B), then to the caregiver (C). We expect the child will follow the robot's gaze cues (D), thereby increasing both the frequency and duration of interaction with their caregiver.

on either the ADOS or ADI-R, alongside a confirmed CBE diagnosis.

5.2.2 Robot-Assisted Intervention System

We describe here the design and expected outcomes of the robot-assisted intervention, the content of the interactions, and the physical and technical components of the system.

Intervention Design. The robot-assisted intervention consisted of 30-minute sessions each day for 30 days and involved triadic opportunities for interaction and shared experiences among the robot, the child, and the caregiver. To achieve this, the robot was designed according to four primary goals: to (i) model realistic social behaviors; (ii) operate autonomously in the home; (iii) deliver personalized content; and (iv) facilitate interactions between the child with ASD and the caregiver.

Intervention Content. The intervention content consists of three interactive games that allowed children with ASD to practice social skills through play. Each of the three games targets one of three social skills: social and emotional understanding, perspective-taking, and ordering and sequencing. The design of these social games is further described in the initial study [3].

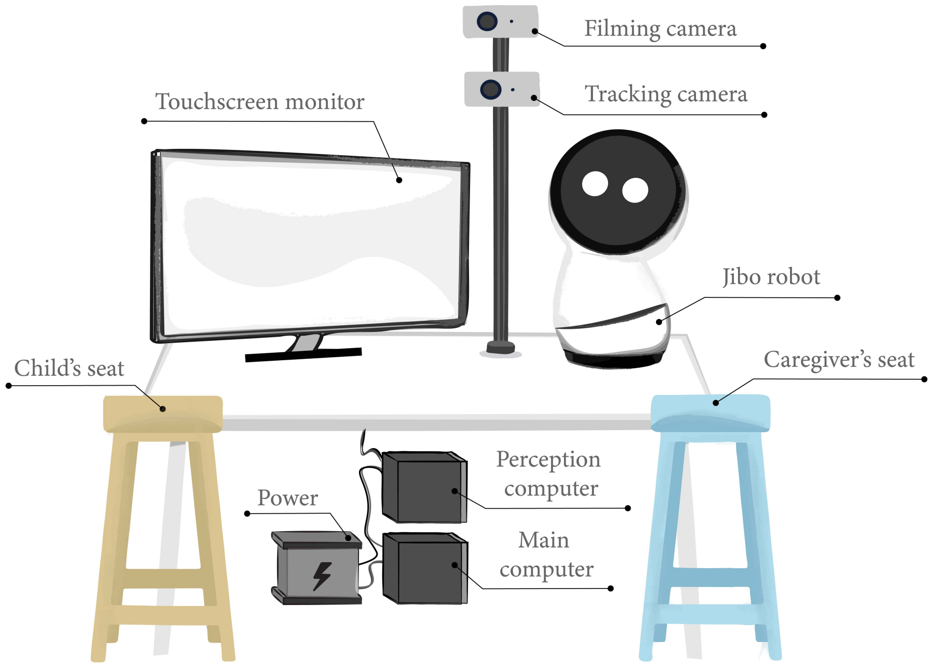


Figure 5.2: System Hardware. The system includes several components to coordinate the robot's behavior, content, and data collection during the intervention sessions.

During the games, the robot demonstrates context-contingent gaze, as illustrated in Figure 5.1. When the child looks at the robot (A), the robot will direct the child's attention to the game content on the screen (B). After, the robot redirects the child's attention to the caregiver (C). We expect the child will follow the robot's gaze (D) and, thus, improve the frequency and duration of their interactions with their caregiver.

System Components. We used the robot Jibo [168] which stands 11 inches tall and has 3 full-revolute axes designed for 360-degree movement. Jibo's onboard capabilities allowed for the programming of personified behaviors such as naturalistic gaze, pose, and movement. Other hardware included a touchscreen monitor, two RGB cameras, a perception computer, and a main computer. The setup is illustrated in Figure 5.2.

Since the robot-assisted intervention relied on modeling appropriate gaze behavior using the robot, Jibo was developed to have a pair of animated eyes as opposed to its default single eye. The perception computer used an elevated camera to track users' attentional focus, relaying this data to the main computer to coordinate the robot's behavior and game content. The touchscreen monitor displayed game content and served as a shared medium between the robot, child, and caregiver. A second camera recorded the sessions for post-study analysis. All components operated within the ROS framework [433].

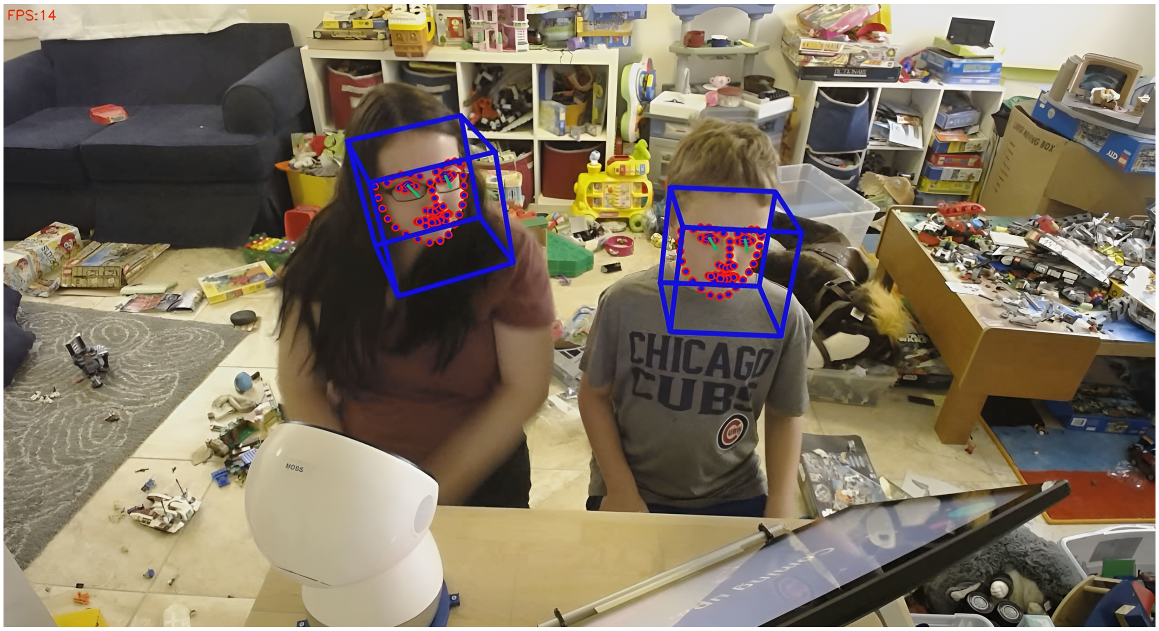


Figure 5.3: Gaze Extraction. We extract the several features such as gaze coordinates and facial landmarks to determine the gaze orientation of the child and their caregiver.

Table 5.1: Detection Accuracy. The performances of the detection algorithm based on manual annotations are shown.

Gaze Component	N	Sensitivity	Specificity	PPV	NPV	AUC	F1
Individual Gaze	9,327	97%	93%	95%	90%	95%	96%
Shared Gaze	6,972	96%	92%	94%	88%	93%	95%
Mutual Gaze	5,823	93%	90%	92%	88%	92%	93%
No Detection	1,195	91%	94%	92%	90%	93%	91%
Overall Performance	23,317	94%	92%	93%	89%	94%	94%

5.2.3 Gaze Extraction

A total of 156 hours of interaction was collected, with each child completing an average of 25 sessions over the month.

Each session recording was pre-processed using OpenFace [434] to extract the gaze orientation of each person in the video feed (Figure 5.3). The resulting features represented information such as the gaze coordinates, facial landmarks, and facial action units for every image frame in the video recordings. Although the caregiver and child sat side-by-side during the intervention, we anticipated natural movement throughout the study, so their locations were not fixed in the analysis. Using OpenFace, we detected multiple faces in each video frame, designating the rightmost as the child

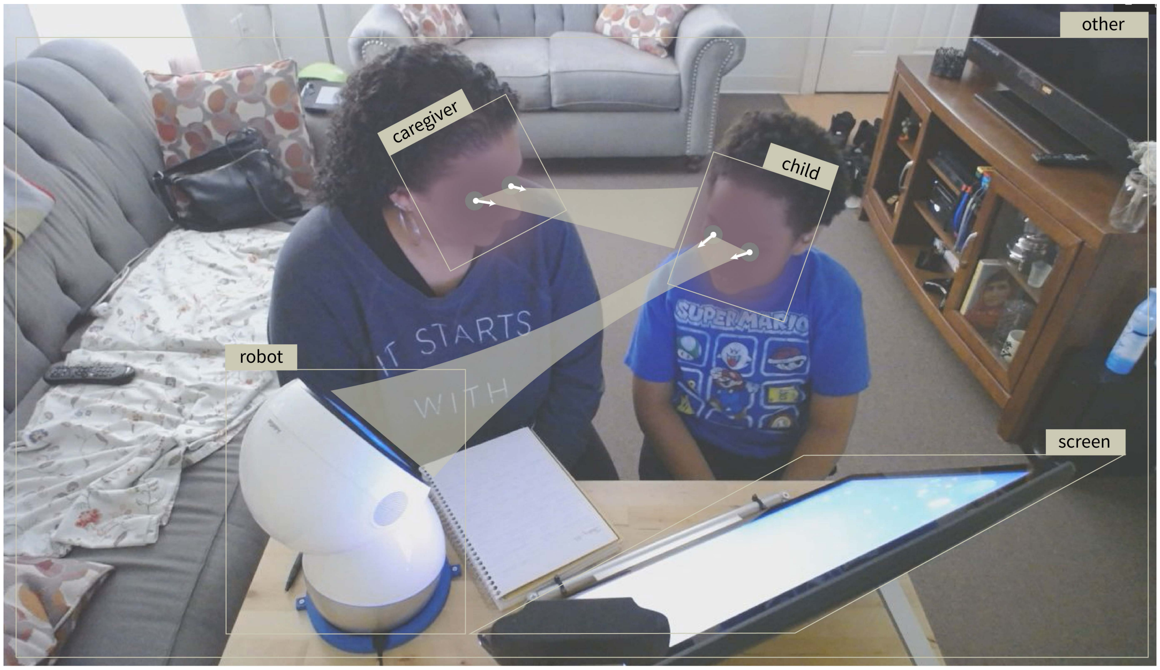


Figure 5.4: Target Detection. Attentional targets are estimated by the intersection of one's visual cone and static object locations.

and the leftmost as the caregiver. If more than two faces appeared—due to other people in the home—the correct faces were manually selected.

To determine the attentional target of the participants, we defined the visual field of the child and caregiver as a cone. Since the location of static primary targets (i.e., screen and robot) are known and that of the caregiver relative to the child can be estimated, a person's gaze is recorded when a target's location falls within their visual cone. For each frame in each video recording, we extract the attentional targets of the child, caregiver, and robot as well as the start time and duration of attention on the target as measured in seconds. Targets include the robot, caregiver, child, screen, and unlabeled objects beyond the intervention content. Gaze data is compressed by grouping consecutive frames where attention remains on the same target, providing event-based data for shifts in attention.

Annotation Method. To assess the accuracy of the gaze detection algorithm, we performed annotations of the participant data. Since we are examining change over time, we randomly selected one session from the first two weeks and another from the last two weeks of each participant's study. 26 sessions were selected across the 13 participants for annotation, representing 12.4 hours or 7.9% of the total dataset. We used the ELAN software [435] to timestamp when the child, caregiver, or robot looked at a target and when they looked away.

To account for the fidelity of human transcriptions, the annotation representing the start of the gaze event was rounded down to the nearest quarter of a second and the annotation representing the end of the gaze event was rounded up to the nearest quarter of a second. As a result, $5,635$ total gaze events were annotated. We aligned all annotated events with the 23,317 detected by the algorithm based on video timestamps to measure overlap. The percentage overlap between detected and annotated events represents the algorithm's accuracy.

Performance by Gaze Component. We identified three primary components of gaze behavior: (i) *individual gaze*, where one person shifted attention to a target, (ii) *shared gaze*, where two or more people looked at the same target, and (iii) *mutual gaze*, where two people made eye contact. An event was labeled "no detection" when it could not be classified, such as when one's eyes were obscured or out of the camera's view.

We evaluated detection accuracy using several standard classification metrics, summarized in Table 5.1. Sensitivity (also known as recall) measures the proportion of actual positives correctly identified, and was high for all three gaze types ($\geq 91\%$). Positive Predictive Value (PPV) and Negative Predictive Value (NPV) indicate the likelihood that positive and negative predictions are correct, respectively. The area under the ROC curve (AUC) provides an aggregate measure of performance across all classification thresholds. F1 scores, which represent the harmonic mean of precision and recall, offer a balanced summary of performance. Finally, weighted averages were used to provide an overall performance summary across all gaze components.

Performance by Subject. We assessed the performance of the detection algorithm in capturing the gaze behavior of both children and their caregivers. This evaluation provided essential insights into the algorithm's reliability across these two user groups in diverse home settings. We employed accuracy as the primary metric to assess whether algorithm correctly identified the attended target for each gaze event.

The detection algorithm demonstrated strong performance in characterizing the gaze behavior of caregivers ($M = 94\%$, $SD = 3.7\%$, $N = 10,909$) and children ($M = 88\%$, $SD = 7.3\%$, $N = 12,408$). Notably, the algorithm yielded a significant difference in average accuracy between caregiver and child data, as determined by a one-tailed z-test for sample proportions ($z = 16.6$, $p \leq 0.001$). This finding indicated that, on average, the algorithm more accurately detected the caregivers' gaze than the children's. Furthermore, an analysis of variance yielded a main effect of the individual, $F = 24.7$, $p \leq 0.001$, indicating that there is a significant difference in the algorithm's performance between caregivers and children.

Figure 5.5: Challenges to Accurate Gaze Detection. In-home environments are inherently unstructured, cluttered, and dynamic, posing several challenges for reliable gaze estimation. These include: (top left) partial occlusions of the child's face; (top right) the presence of toys, siblings, and other family members; (bottom left) non-human faces such as dolls or pets; and (bottom right) frequent motion, especially from the children themselves.

In light of this, we investigated whether specific behaviors contributed to this accuracy difference. we observed a significant difference for gaze events in which a caregiver looked at their child versus not (94%, $N = 1,522$, $z = 3.7$, $p \leq 0.001$) and a significant difference for gaze events in which a child looked at their caregiver versus not (90%, $N = 637$, $z = 3.7$, $p \leq 0.001$) regardless of whether the individual was engaging in independent gaze, shared attention, or mutual gaze.[2] Altogether, this suggested that looking at a target to the immediate right or left significantly influences the accuracy of detection. This is to be expected as turning one's head decreases the amount of facial data to determine gaze. Yet, a significant majority (57%, $z = 28.2$, $p \leq 0.001$) of the events that could not be categorized and received a label of "no detection" by the detection algorithm were of children data. We suspected that this may be because the children showed more physical movement throughout the study than did the caregivers, as confirmed by examining the session recordings.

[2]While these results are significant, we acknowledge that additional statistical refinement—such as the use of hierarchical mixed-effects models—could provide greater clarity by accounting for repeated measures and nested data structures. We report here only the analysis that appears in the published paper [35].

The difference in the algorithm's accuracy for gaze behavior between the caregiver and child is unsurprising, as most open-source datasets for automatic facial behavior analysis focus on neurotypical adults. The algorithm relies on OpenFace's eye gaze estimation [231], which was evaluated by its authors using the MPIIGaze dataset [436], collected from 15 neurotypical adults during everyday laptop use. Despite growing interest in automatic gaze estimation, these methods have not been tested with individuals with ASD or children. Future research should investigate whether OpenFace and similar models are good surrogates for the behavioral annotation of these populations.

Lastly, no significant variations in the algorithm's accuracy were observed when comparing the initial and final stages of the study, by week, or across other categories of gaze behavior.

5.2.4 Dataset

A total of $269,278$ gaze events resulted from this detection analysis. This dataset thus describes gaze behaviors by the frequency and duration an individual engages in throughout their intervention. The current distribution of gaze duration is unimodal and positively skewed. Therefore, a log transform was applied to the duration of gaze to better conform the final dataset to normality, assessed using the Shapiro-Wilk test.

With the resulting dataset, we explore three main components of gaze behavior: (i) *overall gaze* describing general attentional shifts to a target, (ii) *mutual gaze* describing eye contact among the child, caregiver, and robot, and (iii) *joint attention* between the child, caregiver, and robot in which two interacting partners first engage in eye contact, then one partner shifts their gaze to an object, causing their partner to orient their gaze to the same object. We describe these behaviors by the frequency and duration the child or caregiver engages in them over the course of the intervention.

5.3 Results

For each component of gaze behavior, we calculated the averages and variances of gaze instances and duration. We also conducted multiple linear regression analyses to identify predictors of gaze instances for each attentional target. Similar models were used to determine predictors of gaze duration on each target and to assess the moderating effects of clinical measures, including the ADOS, ADI-R, and DAS-II.

We assessed the weekly effects on each gaze component while acknowledging that tasks varied day-to-day based on the children's interests and selections. Because the intervention system adapted to individual preferences, direct comparisons between children or per session were not feasible. However, at the weekly level, each participant was sufficiently exposed to the intervention, despite variations in the daily games and interaction content. Thus, we assessed behavioral changes across participants on a weekly basis throughout the intervention.

We briefly considered whether variability in session length might account for any observable behavioral trends. Over the course of the month-long intervention, participants consistently engaged with the robot for similar durations: an average of 28 minutes (SD = 1.2) during the first five sessions and 27 minutes (SD = 2.5) during the final five sessions. Given this stability and low variability, session length was not included as a factor in our subsequent behavioral analyses.

We also examined whether changes in engagement over time could have influenced gaze behavior. In the original study, caregivers rated daily how easy it was to engage their child in the robot-assisted session. To assess whether sustained engagement reflected genuine interest rather than mere adherence to protocol, Scassellati et al. modeled these caregiver ratings using a cumulative link mixed model with an adaptive Gauss-Hermite quadrature approximation, treating day as a fixed effect and participant as a random effect. The model found no significant effect of day on engagement ratings ($p = 0.82$), indicating that child engagement remained stable throughout the study. Therefore, engagement over time was also excluded as a covariate in our analysis.

5.3.1 Overall Gaze Behavior of the Child

We first investigate the children's distribution of attention across the various targets over time. A multiple linear regression was calculated to predict the log duration of the children's gaze on each attentional target. The regression reveals a significant effect of the week ($F = 19.5$, $p \leq 0.001$) when the target is the caregiver ($\beta = -0.63$, $p \leq 0.001$), when the target is the robot ($\beta = -0.29$, $p \leq 0.001$), when the target is the screen ($\beta = 0.34$, $p \leq 0.001$), and when the target is other than these predefined targets ($\beta = -0.37$, $p \leq 0.001$). Estimated coefficients are denoted as β. A regression did not reveal any significant effects of the clinical measures (ADOS, ADI-R, or DAS-II) on the children's distribution of gaze.

A Tukey's HSD reveals that the average duration of gaze occurrences with the robot ($M = 6.31$ seconds, $SD = 1.08$ seconds; hereafter abbreviated as s) and caregiver ($M = 4.07s$, $SD = 1.21s$) is significantly lower ($p \leq 0.001$) than that with the screen ($M = 70.8s$, $SD = 2.66s$) or targets outside of the intervention ($M = 13.2s$, $SD = 1.40s$). Children's focus on the screen is expected, given that the game's content is a core part of the intervention. Thus, we further examine attention patterns for each target.

Gaze with the Caregiver

A paired t-test performed on the average number of attentional shifts toward the caregiver per week reveals a significant increase in the frequency a child shifts attention toward the caregiver between the first and the last week of the intervention ($t = 3.38$, $p = 0.005$).

A multiple linear regression calculated to predict the log duration of the child's gaze on the caregiver revealed a significant effect of the week ($F = 9.71$, $p \leq 0.001$). The post-hoc analysis reveals a significant decrease in gaze duration until the third week ($M = 2.82s$, $SD = 0.54s$, $\beta = -0.26$, $p \leq 0.001$), where the first two weeks of the study resulted in a significant decrease in gaze duration ($\Delta M = -1.82s$, $p = 0.003$) and a significant decrease in gaze duration between the second and third week ($\Delta M = -1.45s$, $p = 0.006$). However, in the last week of the study, we observed a significant increase in the gaze duration of the child with the caregiver ($\Delta M = 1.86s$, $p \leq 0.001$). This change indicates that increased gaze duration with the caregiver occurred after having engaged with the intervention for at least three weeks.

The multivariate linear regression showed no significant effect of clinical scores on children's gaze toward the caregiver. Overall, participants with ASD increased both their gaze frequency and duration toward the caregiver over time.

Gaze with the Robot

A paired t-test performed on the average number of attentional shifts towards the robot per week reveals a significant increase in the frequency a child with ASD shifts gaze to the robot beginning in the third week of the intervention ($t = 2.65$, $p = 0.03$). However, this significant increase does not persist into the last week of the intervention ($t = 1.34$, $p = 0.21$). This change indicates that participants with ASD showed an increased tendency of looking at the robot only after two weeks of the intervention.

A regression calculated to predict the log duration of the child's gaze on the robot reveals a significant effect of the week ($F = 18.8$, $p \leq 0.001$). The post-hoc analysis reveals a significant decrease in gaze duration with the robot throughout the

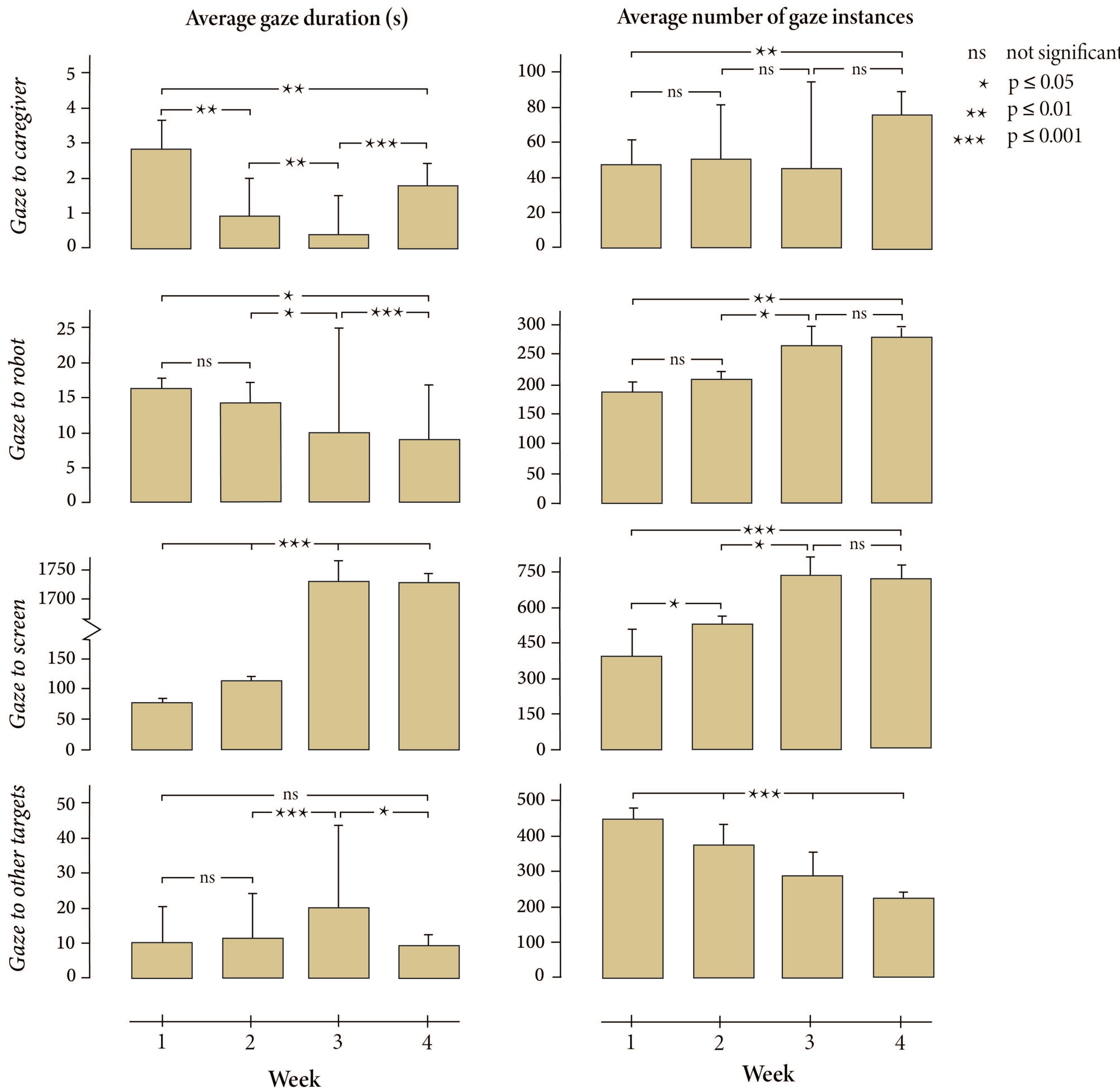


Figure 5.6: Children's Average Gaze Duration & Frequency Per Week. The change in children's average gaze duration (on the left) and gaze instances (on the right) with each attentional target per intervention week are shown. The intervention led to significant increases in both the amount and distribution of children's gaze directed toward their caregivers, as compared to other targets within and outside the intervention setting. We also observe notable week-by-week variation, with clearer improvements in gaze behavior emerging after the second week.

study. The decrease becomes significant in the third week of the study ($M = 10.0s$, $SD = 15.3s$, $\beta = -0.17$, $p \leq 0.05$) as compared to the previous week ($M = 14.5s$, $SD = 5.14s$, $\beta = -0.07$). This rate of decrease persisted to the end of the study ($\beta = -0.17$, $p \leq 0.001$) and therefore suggests that children with ASD were more likely to shift attention away from the robot over time.

However, a Levene variance test shows log gaze durations varied significantly by week ($w = 7.96$, $p \leq 0.001$). A significant change in the gaze duration variance in children with ASD occurred until two weeks into the study and persisted until the end ($p \leq 0.001$). This change suggests that, although the rate of decreased attention to the robot was similar among all participants with ASD, the variability in gaze duration among participants with ASD was greater later in the intervention.

In addition, the regression showed a significant effect of the ADOS severity score ($\beta = 1.21$, $p = 0.006$) and all categories of the ADI-R (reciprocal social interactions, $\beta = 0.59$, $p = 0.007$; communication, $\beta = -0.35$, $p = 0.007$; restricted, repetitive, and stereotyped behaviors, $\beta = -0.87$, $p = 0.005$; history of early abnormal development, $\beta = 1.32$, $p = 0.006$), and of the DAS-II (verbal reasoning, $\beta = 0.02$, $p = 0.02$; nonverbal reasoning $\beta = 0.52$, $p = 0.006$; spatial reasoning, $\beta = 0.22$, $p = 0.007$; general conceptual ability or GCA, $\beta = -0.45$, $p = 0.007$). Children with higher ASD severity scores, lower communicative ability, or more stereotyped behaviors were more likely to show increased attention toward the robot.

Gaze with the Screen

A paired t-test performed on the number of attentional shifts reveals a significant increase in the frequency the children look at the screen between the first and last week of the intervention ($t = 5.50$, $p \leq 0.001$).

A regression calculated to predict the log duration of the child's gaze on the screen revealed a significant effect of the week ($F = 76.8$, $p \leq 0.001$). The post-hoc analysis reveals a significant increase in gaze duration with the screen throughout the study ($\beta = 0.33$, $p \leq 0.001$), suggesting that children with ASD consistently attend longer to the screen over time.

However, a Levene variance test shows log gaze duration significantly varies among children with ASD by week ($w = 24.43$, $p \leq 0.001$). This significant change begins in the third week of the study ($M = 1737.8s$, $SD = 33.7s$), as compared to the previous week ($M = 100.0s$, $SD = 5.92s$), and persists to the end of the study. This suggests that, although the rate of increased attention to the screen is similar among all participants with ASD across each week, the variability in gaze duration is greater after two weeks into the intervention.

The effect of clinical scores[3] on gaze duration with the screen, although similar in magnitude, is in the opposite direction of that with the robot; children with lower severity and stereotyped behaviors, and higher communicative ability and IQ showed increasing attention toward the screen.

Gaze with Other Targets

A paired t-test reveals a significant decrease in the frequency in which a participant with ASD turns attention to targets outside of the intervention between the first and last week ($t = 4.32$, $p \leq 0.001$).

A multiple linear regression calculated to predict the log duration of the child's gaze outside of the intervention's targets revealed a significant effect of the week ($F = 7.06$, $p \leq 0.001$). The post-hoc analysis reveals a significant increase in gaze duration with external targets after the second week ($\beta = 0.05$, $p \leq 0.001$), but this significant increase is not observed throughout the study ($\beta = -0.01$, $p = 0.82$). This transition after the second week ($M = 10.72s$, $SD = 12.26s$) is reflected in a significant increase in average gaze duration with external objects between the third week ($M = 19.1s$, $SD = 25.0s$, $p \leq 0.001$), and significant decrease the fourth week ($M = 7.08s$, $SD = 4.37s$, $p = 0.013$). The differences of variance between the weeks is also significant ($w = 13.41$, $p = 0.004$) after the second week and supports previous results indicating that the behavioral variability among those with ASD was greater after two weeks into the intervention.

The effect of clinical scores[4] on gaze duration is similar in both magnitude and direction as that with the robot; children with lower ASD severity scores, high communicative ability, or less stereotyped behaviors are more likely to show increased attention toward the robot.

[3] ADOS calibrated severity score ($\beta = -1.04$, $p = 0.002$); all ADI-R categories: reciprocal social interactions ($\beta = -0.56$, $p = 0.001$), communication ($\beta = 0.36$, $p \leq 0.001$), restricted, repetitive, and stereotyped behaviors ($\beta = 0.76$, $p \leq 0.001$), history of early abnormal development ($\beta = -1.14$, $p = 0.002$); and all DAS-II categories: verbal reasoning ($\beta = 0.02$, $p = 0.01$), nonverbal reasoning ($\beta = -0.45$, $p = 0.002$), spatial reasoning ($\beta = -0.19$, $p = 0.002$), GCA ($\beta = 0.39$, $p = 0.002$).

[4] ADOS calibrated severity score ($\beta = 1.26$, $p \leq 0.001$); all ADI-R categories: reciprocal social interactions ($\beta = 0.66$, $p \leq 0.001$), communication ($\beta = -0.41$, $p \leq 0.001$), restricted, repetitive, and stereotyped behaviors ($\beta = -0.91$, $p \leq 0.001$), history of early abnormal development ($\beta = 1.39$, $p \leq 0.001$); and all DAS-II categories: verbal reasoning ($\beta = 0.01$, $p = 0.04$), nonverbal reasoning ($\beta = 0.55$, $p \leq 0.001$), spatial reasoning ($\beta = 0.23$, $p \leq 0.001$), GCA ($\beta = -0.46$, $p \leq 0.001$).

5.3.2 Overall Gaze Behavior of the Caregiver

Using a paired t-test, we found a significant increase in caregiver gaze shifts to the robot from the first to the last week of the intervention ($t = 7.97$, $p \leq 0.001$), with no significant change in gaze to the screen ($t = 1.33$, $p = 0.21$). Caregivers showed a significant decrease in gaze toward the child over the study period ($t = -15.2$, $p \leq 0.001$) and a significant increase in shifts to external targets ($t = -3.82$, $p = 0.002$).

A regression of the log duration of the caregiver's gaze also reveals a significant effect of the week ($F = 70.7$, $p \leq 0.001$) and when the target is the child ($\beta = -0.23$, $p \leq 0.001$), robot ($\beta = 0.16$, $p \leq 0.001$), screen ($\beta = 0.67$, $p \leq 0.001$), or outside of these predefined targets ($\beta = 0.22$, $p \leq 0.001$).

A Tukey HSD revealed that caregivers spent significantly more time attending to the screen ($M = 166.0s$, $SD = 3.95s$) compared to other targets, including those outside the interaction ($M = 32.4s$, $SD = 3.36s$), the robot ($M = 14.1s$, $SD = 2.32s$, $p = 0.03$), and their child ($M = 6.03s$, $SD = 1.32s$, $p \leq 0.001$). The increased attention to the robot, screen, and other targets beyond the intervention indicates a quicker shift in focus away from the child. Although children's gaze duration and frequency toward their caregiver increased significantly (though not consistently across the weeks, see Section 5.3.1), caregivers' gaze toward their child decreased in a more pronounced, consistent manner over time. Figure 5.8 shows the magnitude of change for the children and caregivers.

The regression calculated to predict the log duration of the caregiver's overall gaze reveals significant effects of their child's ADOS calibrated severity score ($\beta = 1.00$, $p \leq 0.001$), all categories of the ADI-R (reciprocal social interactions, $\beta = 0.46$, $p \leq 0.001$; communication, $\beta = -0.28$, $p \leq 0.001$; restricted, repetitive, and stereotyped behaviors, $\beta = -0.73$, $p \leq 0.001$; history of early abnormal development, $\beta = 1.06$, $p \leq 0.001$), and all categories of the DAS-II (verbal reasoning, $\beta = 0.02$, $p \leq 0.001$; nonverbal reasoning, $\beta = 0.42$, $p \leq 0.001$; spatial reasoning, $\beta = 0.19$, $p \leq 0.001$, GCA, $\beta = -0.37$, $p \leq 0.001$). Further investigation reveals that the significant effects of clinical measures occur only when caregivers focused on the robot or child. This suggests that caregivers of children with high ASD severity scores engaged in longer gaze behavior with both the robot and child over time. Moreover, the direction of these effects is similar for both caregivers and children: both engage in longer gaze with the robot when clinical measures indicate high ASD severity, low communicative ability, or more stereotyped behaviors.

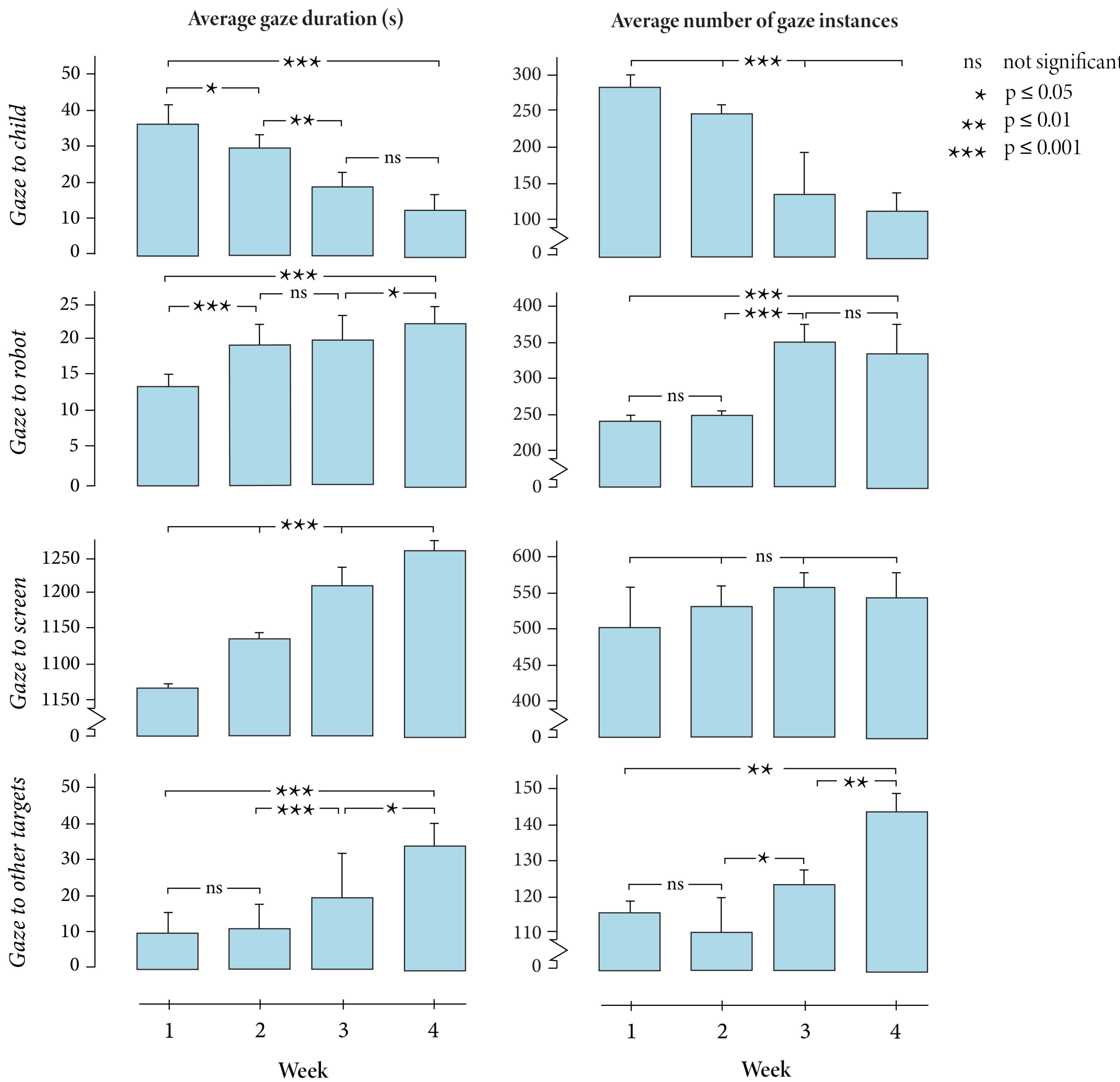


Figure 5.7: Caregivers' Average Gaze Duration & Frequency Per Week. This figure illustrates weekly changes in caregivers' average gaze duration (left) and gaze frequency (right) toward various attentional targets during the intervention period.

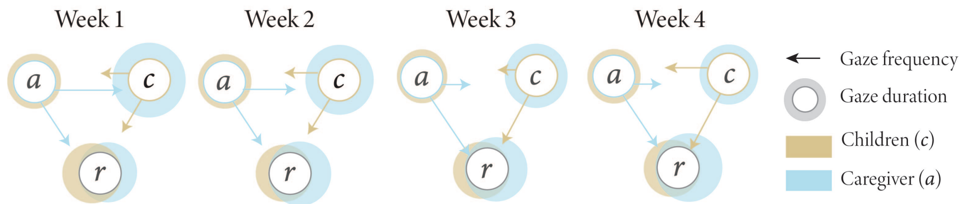


Figure 5.8: Change by Week. Average gaze duration and frequency for adult caregivers (a) and children (c) are shown. Circle diameters indicate average duration (in seconds) toward each other and robot (r), while line lengths indicate frequency. This summarizes the bar chart representations for children and caregivers, shown in Figures 5.6 and 5.7, respectively.

5.3.3 Joint Attention Based on Mutual Gaze

The previous analyses focused on trends in individual gaze instances and durations, as well as predictors such as weekly exposure to the intervention and clinical measures. We now expand our scope to examine the contingency of gaze among the robot, children, and caregivers throughout the intervention. We define this contingency through *joint attention involving eye contact*, which occurs when two individuals engage in mutual gaze, and one shifts their gaze to an object, prompting the other to follow. This gaze following reflects an expectation-based orienting, where one person's change in gaze cues the other's attention [437]. It is anticipated that joint attention initiated by mutual gaze leads to greater motivation to follow gaze cues and longer durations of shared gaze [438, 439].

Between Child and Caregiver

A paired t-test revealed a significant increase in spontaneous mutual gaze between children and caregivers ($t = 4.31$, $p = 0.009$). A regression predicting the duration of shared gaze following joint attention showed a significant effect of the week ($F = 10.30$, $p \leq 0.001$), with a marked increase in the first week that persisted throughout the study ($\beta = 0.27$, $p \leq 0.001$). These findings suggest that joint attention between children and caregivers increased, with more frequent mutual gaze leading to longer periods of shared gaze over time. No significant effects of clinical measures were found in the regression.

We also observed a significant decrease in the duration of mutual gaze from the first to the last week of the study ($\Delta M = -1.43s$, $\beta = -0.20$, $p = 0.002$). Despite the significant increase in joint attention and the resulting shared gaze between the child and caregiver, the duration of mutual gaze decreased. This may be viewed as a positive outcome, as shorter durations of the joint attention cue (i.e., eye contact) allows for longer durations of shared attention.

Between Robot and Child

A paired t-test on the frequency of the child's gaze toward their caregiver following eye contact with the robot indicates a significant increase after the second week of the intervention ($t = 4.56$, $p \leq 0.001$). Similarly, after the second week, when the child shifted gaze away from the caregiver, they more frequently redirected their attention back to the robot ($t = 3.46$, $p = 0.004$). A regression analysis also shows a significant effect of the week ($F = 8.44$, $p \leq 0.001$). Joint attention between the robot and child significantly increased from the first week and continued through the study ($\beta = 0.03$, $p = 0.006$).

We observed a similar trend in the duration of eye contact between the child and the robot ($\beta = 0.05$, $p \leq 0.001$) and in gaze duration when the joint attentional target is the caregiver ($\beta = 0.27$, $p = 0.05$). The increase in joint attention indicates that the children directed more attention to their caregiver while engaging with the robot over time. The regression, however, does not show a significant effect of clinical scores.

Between Robot and Caregiver

A paired t-test of the frequency of gaze of the caregiver towards their child following eye contact between the robot and the caregiver by week suggests that the caregiver significantly shifted gaze more often toward the child after following the gaze of the robot throughout the study ($t = 2.80$, $p = 0.02$). When shifting gaze away from the child, caregivers shifted their attention more often to the screen ($t = 10.1$, $p = 0.004$).

A regression predicting instances of joint attention between the caregiver and robot shows a significant weekly effect ($F = 5.00$, $p = 0.002$), with a notable increase starting after the second week and persisting throughout the intervention ($\beta = 0.07$, $p = 0.004$). Additionally, we observed a significant increase in the duration of eye contact between the caregiver and robot ($\beta = 0.33$, $p \leq 0.001$), as well as in gaze duration when the child is the joint attentional target ($\beta = 0.07$, $p = 0.006$), but not when the target is the screen ($\beta = -0.05$, $p = 0.52$). This suggests that caregivers increasingly focused on their children while engaging with the robot over time.

The effects of a child's clinical severity on their caregiver's overall gaze[5] and the

[5] ADOS calibrated severity score ($\beta = 1.38$, $p \leq 0.001$); ADI-R categories: reciprocal social

joint attention between the caregiver and robot are similar in both direction and magnitude; joint attention between the caregiver and robot increased when their child exhibits higher ASD severity, lower communicative ability, or more stereotyped behaviors.

5.4 Discussion

Scassellati et al. [3] introduced a robot-assisted intervention system that provided personalized, on-demand cognitive and social support for children with ASD. We expand the definitions, detection methods, and analysis of user behavior to better capture the effects of a long-term, in-home social robot intervention for ASD. Our findings center on three key themes: (i) the intervention improved gaze behavior in children with ASD; (ii) behavioral variability among participants increased significantly after two weeks of engagement; and (iii) diagnostic measures like the ADI-R, ADOS, and DAS-II proved to be strong predictors of behavioral change for both caregivers and children. These insights are crucial for designing effective robot-assisted social skills interventions and understanding behavioral trends in ASD.

5.4.1 Improvements in Gaze Behavior

The social robot promoted appropriate gaze behavior during the intervention, leading to improved spontaneous gaze between children with ASD and their caregivers. Children were significantly more likely to direct their attention and make eye contact with their caregivers. Our analysis revealed significant increases in instances of joint attention, spontaneous mutual gaze, and the duration of shared gaze between the pairs. However, while children engaged in eye contact with their caregivers more frequently, the duration of eye contact prior to shared gaze decreased over time. We view this as a positive outcome, as it indicates that children needed less time for the joint attentional cue (eye contact) to maintain longer periods of shared attention with their caregivers.

Furthermore, the children's gaze with the caregiver was contingent on that of the robot throughout the study: children with ASD were more likely to engage in longer eye contact with their caregiver after they saw the robot shift its attention to the

interactions ($\beta = 0.70$, $p \leq 0.001$); communication ($\beta = -0.43$, $p \leq 0.001$); restricted, repetitive, and stereotyped behaviors ($\beta = -0.99$, $p \leq 0.001$); history of early abnormal development ($\beta = 1.49$, $p \leq 0.001$); DAS-II categories: verbal reasoning ($\beta = 0.03$, $p \leq 0.001$); nonverbal reasoning ($\beta = 0.61$, $p \leq 0.001$); spatial reasoning ($\beta = 0.26$, $p \leq 0.001$); GCA ($\beta = -0.53$, $p \leq 0.001$).

caregiver. This contingency of gaze is also true of caregivers: caregivers were more likely to engage in longer eye contact with their child after they saw the robot shift its attention to the child. This suggests that a robot designed to redirect a person's attention by modeling the shift in gaze may be effective at improving the frequency of eye contact.

Ultimately, gaze following with the robot was natural, increased throughout the study for both children and caregivers, and encouraged more frequent eye contact and shared attention between the children and caregivers. Using a joint attention probe, Scassellati et al. [3] found significant improvements in joint attention among children with ASD following a robot intervention. Similarly, this study confirms consistent joint attention gains with both the caregiver and the robot. Yet, we must acknowledge that the impact of the robot or any other system component cannot be measured independently. Furthermore, the sustainability of the observed gains may depend on ongoing participation in the intervention or additional support. These improvements were noted during the intervention, but further research is necessary to determine whether they persist long after the study concludes. We present this as a limitation of this study and an area for future work.

5.4.2 Timing & Variability of Skill Improvements

Several improvements in gaze behavior emerged *only* after two weeks into the intervention. For instance, joint attention between the robot and caregivers significantly increased only after the second week. Additionally, children's gaze duration toward their caregiver initially decreased during the first two weeks, followed by a significant increase in the final two weeks. Hence, we recommend that similar social skills interventions be evaluated over a duration longer than two weeks to better capture the potential for significant behavioral change.

It is well known that individuals with ASD show a broad spectrum of challenges and (dis)abilities, and vary greatly in their levels of social functioning. Although the participants were high-functioning individuals with ASD and able to understand the intervention's content, we observed significant variability in the gaze behaviors between users *only* after two weeks. This variability among the children was especially evident in gaze towards objects that were initially novel: the screen and the robot. While each child's gaze behavior with these objects followed a similar pattern for the first two weeks, their behaviors with these objects diverged significantly after. Based on these findings, we recommend that interventions aimed at improving gaze behaviors in children with ASD be evaluated for more than two weeks, allowing for novelty effects to subside and increased individual variability to emerge.

5.4.3 Predictive Power of Diagnostic Measures

Scassellati et al.'s joint attention probe found that children with lower nonverbal ability, as measured by the DAS-II, showed greater gains in joint attention skills. Our analysis further supports this, revealing a strong positive relationship between nonverbal ability and joint attention, suggesting that children with lower nonverbal reasoning had more capacity to grow in terms of joint attention skills.

Furthermore, the strong correlation between clinical measures of ASD severity and gaze behaviors suggests that these metrics can be valuable for predicting intervention outcomes. For example, children with higher ASD severity and lower nonverbal ability showed increased attention to the robot while their attention to the screen decreased over time. These scores not only predicted the children's behaviors but also their caregivers'. Caregivers of children with high ASD severity engaged in longer gaze interactions with both the robot and their child. Being able to anticipate user outcomes based on clinical severity can influence how we think about the intervention's effectiveness and allow researchers to streamline the process by reducing the need for constant clinician oversight.

5.4.4 Implications & Limitations

We recommend that designers of social skills interventions for ASD leverage these findings by recognizing the potential of robots to foster appropriate gaze behavior among users. Additionally, the results indicate that such an intervention on gaze behavior should be evaluated for at least two weeks to account for the decline of novelty effects and the subsequent behavioral variability among individual users. Further research is necessary to determine how effectively clinical assessments of ASD predict the outcomes of robot-assisted social skills interventions. The strong correlation between clinical scores and gaze behavior suggests that these assessments could reliably predict the behaviors of both children with ASD and their caregivers during the intervention. This relationship may reduce the need for constant oversight by a clinician or for disrupting in-home interactions to administer tests that may not fully capture the specific skills targeted by the intervention.

While our study provides valuable insights into gaze behavior during a long-term, in-home social robot intervention for ASD, it has several limitations. The small sample size (13 children with ASD and 13 caregivers) limits the generalizability of our findings. Behavioral changes from training often require weeks or months, and while the month-long intervention captured novelty effects and early impacts, longer studies are needed to assess the sustainability of improvements. Future research with larger samples and extended interventions is essential to better understand the long-term effects of social robot interventions on gaze behavior in ASD.

5.5 Summary

This chapter examined behavioral change resulting from a month-long, in-home social robot intervention for individuals with ASD. The findings offer design recommendations for developing clinically meaningful SAR-based social skills interventions and provide deeper insights into the behavioral patterns and learning trajectories associated with ASD. We demonstrate that the robot-assisted intervention significantly improved multiple aspects of gaze behavior in children with ASD, with notable individual variation in the timing and trajectory of these improvements. Importantly, early diagnostic measures were strong predictors of long-term gaze behavior for both children and their caregivers. Together, these findings advance our understanding of behavioral patterns in ASD and underscore the clinical potential of robot-based interventions.

We continue our investigation of robot-assisted interventions aimed at enhancing specific social skills in individuals with ASD. While the landmark study by Scassellati et al. [20] analyzed here marked the first in-home SAR intervention for children with ASD, the following chapter presents the first in-home SAR intervention designed specifically for *adults* with ASD.

Chapter 6

A Social Robot for Improving Interruptions Tolerance and Employability in Adults with ASD

Despite decades of progress in ASD research, the vast majority of studies and clinical programs have focused almost exclusively on children. Although social, emotional, and functional challenges are well-documented to persist, and in some cases intensify, in adulthood, relatively few studies have addressed how to support adults with ASD across life transitions. In this chapter, we explore how robots can support employability and workplace readiness for adults with ASD. We developed a robot-directed intervention that simulated common workplace encounters, promoting role-play and naturalistic social practice while integrating into participants' daily home routines. Over the course of a week, users engaged in managing unexpected social demands and developed strategies for cognitive and attentional regulation. Behavioral data and participant feedback revealed increased resilience to work-relevant interruptions, positive perceptions of the robot's usefulness for supporting employment goals, and early evidence of skill generalization beyond the specific HRI. This study[1] represents the first in-home, robotic intervention designed for adults with ASD.

6.1 Introduction

Individuals with Autism Spectrum Disorder (ASD) exhibit social skill deficits such as difficulties with reciprocal social interaction, interpersonal communication, and insistence on behavioral and environmental sameness [194]. These individuals show a broad spectrum of challenges and (dis)abilities, and vary greatly in their levels of social functioning [440].

[1]This chapter is adapted from our published work: **Ramnauth, R.**, Adéníran, E., Adamson, T., Lewkowicz, M. A., Giridharan, R., Reiner, C., & Scassellati, B. (2022, March). A Social Robot for Improving Interruptions Tolerance and Employability in Adults with ASD. In the *2022 17th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 4-13). IEEE. [59].

Figure 6.1: Robot-Assisted Interruptions Training in the Home. The Interruptions Skills Training and Assessment Robot (ISTAR) is designed to help adults with ASD practice handling interruptions in their home, therefore providing workplace-relevant skills training in an intuitive and organic way. The collage on the left illustrates typical interactions between the system and an adult with ASD in four home deployments. The rightmost image shows the system in a user's home.

ASD is a costly condition—both economically and in terms of human experience. Approximately 85% of adults with ASD face chronic unemployment or underemployment [441], a significantly higher rate than that observed in adults with other developmental disabilities [442]. Creating an inclusive workplace by improving the employability of adults with ASD would result in financial independence and higher quality of life for the individual. Furthermore, many individuals with ASD have unique strengths and abilities, such as attention to detail, task persistence, and strong work ethic—skills highly valued across various employment sectors yet often underutilized [443, 444].

Finding and maintaining employment is complex and involves several stages from submitting a job application or participating in an interview, to navigating the responsibilities and expectations once employed. Each stage demands an ability to adapt to unforeseen circumstances and recover effectively from interruptions. Empirical research demonstrates that commonplace interruptions can result in significant lost work, costly errors, or safety violations [445, 446]. We are motivated to study interruptions due to their frequency in the workplace [447] and their measurable effect on workflow [448].

Unfortunately, existing methods to mitigate the effect of interruptions focus on restructuring the workplace environment to limit the frequency of interruptions [449]. Understanding how interruptions impact current workflows, characterizing an individual's capacity to regulate attention effectively between tasks, and training individuals to support better error-free interruptions recovery are desirable for any person that experiences interruptions.

However, interruptions can be especially challenging for people with ASD. Workplace distractions, unpredictability, and uncertainty may pose heightened challenges for individuals with ASD, given their characteristic social and communicative difficulties [450]. Aaron Likens, an Easterseals national representative and adult with ASD, reports: “That's the way my brain is; once at speed I can focus with perfect clarity but that one interruption can bring about a complete change in ability to focus or achieve a task, hence why the unsuspecting interrupter is going to get what sounds like an angry answer.” [451]. Individuals with ASD may experience not only the cognitive disruption, but also the social-emotional consequences of an interruption more acutely than others [450].

Social robotics has the potential to address the critical gap of job-relevant interruptions training for this unique and understudied population [452, 453]. Compared to other technologies, a robot provides a physical component to the training experience that makes it difficult for users to ignore or silence its prompts for interaction. Furthermore, we consider socially assistive robotics (SARs) because it merges traditional robotics and computational methods to improve access to personalized, socially situated, and physically co-present interactions [454]. In other words, a SAR for social skills training creates a situated, embodied interaction that requires users to engage in socially appropriate ways.

Research has established that SARs for ASD interventions can result in positive and productive outcomes [20]. A recent study indicates that robot-assisted therapy may be effective for improving interruptions tolerance in adults with ASD [452]. Preliminary work find that aspects of face-to-face communication can be supported with robot interactions [455] and in-home robot-led training can be applicable to the workplace [452].

Leveraging this promise of SARs for ASD interventions, we developed the *Interruption Skill Training and Assessment Robot* (ISTAR), an in-home autonomous training system that helps adults with ASD practice handling workplace-relevant interruptions. This system targets social-skills development in a familiar environment and can provide valuable support for adults with ASD as they find and maintain employment.

6.2 Background

In this section, we review recent literature identifying common barriers to gainful employment for adults with ASD, with a particular emphasis on the importance of interruptions training for improving workplace readiness. Finally, we explore the potential of SARs to address the persistent gap in accessible, contextually grounded social skills training for this population.

6.2.1 Job Skills Training for Adults with ASD

Few individuals with ASD have been trained in the vocational skills needed to find and maintain gainful employment. The number of under- and unemployed adults with ASD is exceptionally high, even compared to those in similar disability groups [456]. Most job training for adults with ASD that have been demonstrated to be effective target specific on-the-job tasks such as mail sorting, photocopying, and stocking shelves [457]. Consequently, traditional job training overlooks many of the soft skills essential to job maintenance, including time management, organization, and customer or co-worker interactions. These skills are often the most difficult for persons with ASD.

In all, interventions for ASD do not yet capture the heterogeneity of impairment [440], the demographic [458], or the range of services needed to help adults function with purpose in their communities [459]. Although employment interventions for ASD have been developed, many are not clinically meaningful and lack clear evidence concerning their efficacy [460, 461]. Due to the vast heterogeneity of ASD, a "one size fits all" approach is insufficient and counterproductive [462].

6.2.2 Interruptions Training

It is commonly understood that the more people practice performing a particular task, the better they are able to perform that task (i.e., the practice effect; [463]). It reasonably follows that the more an individual practices with interruptions, the better they will become at recovering from interruptions. Research examining the effects of repeated exposure to interruptions supports this view [464, 465]. Two standard behavioral metrics are used to measure the disruption caused by an interruption and to evaluate the success of interruptions training: interruption lag and resumption lag. *Interruption lag* is the time needed to address an interruption once it has happened. Similarly, *resumption lag* is the time needed to "collect one's thoughts" and resume the original task after an interruption is over [466]. Performing a task while experiencing interruptions over several sessions reduces interruption and resumption lags to improve overall performance [465]. However, the source of improved performance is not yet understood. It remains unclear whether improvement arises from repeated exposure of the primary task alone, from reduced cognitive demand due to the practice effect, from experiencing the co-occurrence of the primary and interrupting tasks, or from a more general learning process where exposure to specific interrupting tasks leads to improvement at handling any interruption [465, 467].

Yet, to minimize the disruptive effects of interruptions, it is not sufficient for people to gain expertise at specific primary tasks [465, 468]. Instead, they must also gain expertise at performing tasks with interruptions. As a result, individuals who work in environments subject to many interruptions benefit from practicing workplace-relevant primary and interrupting task pairs. As it is difficult to account for all possible interruptions when developing an interruptions training platform, both task-analytic and observational techniques must be applied to identify the types of interruptions most prevalent in a given environment. For example, in the safety-critical environment of the flight deck, the most common interruptions are radio contact with air traffic controllers, requests from flight attendants, and alerts from the aircraft itself [465]. Incorporating these common interruptions into flight simulation for pilot training has reduced disruptions on the flight deck where error tolerance is at or near zero percent [469, 470].

Nevertheless, it is an ambitious task to compile a comprehensive and continuously relevant set of task pairs that will manifest in the real-world. Job training programs should incorporate general workplace interruptions into the practice of primary work tasks to ensure that individuals will be able to recover effectively when faced with real-world interruptions.

6.2.3 Social Robotics for ASD Skills Training

Recent evidence suggesting that technology-driven interactions enable better social understanding for adults with ASD [20, 471] has encouraged researchers to explore technology for workplace interventions [452]. Emerging "Inclusion Engineering" efforts [472] create environments where marginalized individuals can master various everyday tasks that are key to productive employment. Virtual environments have been developed to role-play common employment scenarios such as job interviews. These role-playing scenarios have demonstrated long-term post-intervention improvements [473]. Leveraging the advantages of an embodied system [20], human-robot interactions have the potential for effective skills training for improving the employability of adults with ASD.

SARs have been shown to increase both compliance [474] and learning gains [368] in similar applications. Well-grounded evidence increasingly pervades the literature to affirm that interaction between individuals with ASD and embodied artificial agents encourages prosocial behaviors [206], sustains attention, induces spontaneous and appropriate social behavior, decreases stereotyped and repetitive behaviors [475], optimizes cognitive learning gains [476], and enhances social engagement [20, 21]. In all, a robot that engages its users in social-skills training can be a valuable tool for adults with ASD.

6.3 Design Goals

Designing ISTAR was an iterative process. We first examined responses to interviews assessing the state of employment of adults with ASD and the potential for interruptions training. These interviews suggested that an in-home social robot training platform would be applicable to improving users' resiliency to workplace interruptions. We describe here our design goals inspired by the recommendations gathered from these interviews. Later, we improved our prototype based on survey assessments from adults with ASD and employers (Section 6.5). The improvements directly addressed our design goals and made ISTAR more autonomous, robust, and responsive for a home environment. Ultimately, the final system was ready for deployment into homes of adults with ASD (Section 6.6).

Our collaborators [477] conducted individual and focus group interviews with employers, service providers, and adults with ASD to achieve a first-hand account of their perceptions of employment and the current workforce. This interview series included a total of 23 participants. Ten participants were divided into two focus groups, and 13 were interviewed individually. The interviews involved four target groups: adults with ASD, current or potential employers, educational representatives, and service providers. The employers who participated in the study interviews had at least five years of experience hiring and working with adults with ASD. The service providers and educational representatives facilitated cross-talk and liaised between adults with ASD and potential employers. In these interviews, individuals with ASD highlighted that *interruptions* in the workplace from other people were "problematic" and considered a barrier to maintaining employment. Employers reported that successful employees with ASD are part of peer support programs that encourage socialization and role-playing situations as an effective form of preparation. The design requirements of our robot prototype address these insights by providing role-based interruptions training to its users.

In light of this, to improve tolerance to real-world interruptions, the system should provide workplace-relevant interruptions training through role-playing. With efficient and relevant training, we expect users will improve their tolerance for workplace interruptions where, over time, the interruptions will become less disruptive, allowing them to return to their primary task quickly. There are four primary design goals for ISTAR:

1. **Embodied**. The system should be embodied as a robot. A social robot can produce measurable learning outcomes [368], provide a physical component to the training experience that improves compliance [474], and express realistic cues that encourage socially appropriate responses from users [478].

2. **In-the-home**. The system should be designed to provide training in the home. Therefore, users can interact with ISTAR to avoid potential stigma from colleagues, and without needing approval from or declaring a diagnosis to their employers. Although similar systems for studying interruptions [452] have been designed for clinical or laboratory settings where environmental conditions can be controlled or planned for [3], the home is a dynamic, unstructured environment that demands more complex sensing and behavioral decisions.

3. **Autonomous**. Training should be fully autonomous; it should not be necessary for someone with technical expertise to adjust or control the system once it is given to the user.

4. **Realistic**. ISTAR should provide realistic interactions that are appropriate and similar to interruptions that occur in the workplace, respond in real-time, and express human-like behaviors such as naturalistic gaze, movement, and speech.

ISTAR is designed to be an in-home interruptions training robot. Its interaction model emphasizes frequent, brief, and contextually situated prompts intended to capture the user's attention without overwhelming or disengaging the user. This low-burden design allows ISTAR to integrate naturally into the rhythm of users' everyday routines, providing micro-interventions that simulate realistic workplace interruptions. After each interaction, users exercise their resiliency to interruptions by resuming their original activities.

Figure 6.2 illustrates ISTAR delivering an interruption. The user is engaged in his primary task of reading while ISTAR sits on the desk beside the user. ISTAR is configured to initiate an interruption only when its user is within its camera's view. The first frame (A) shows the user focusing on a primary task. In frame B, ISTAR initiates an interruption by asking the user a question to capture his attention. Then, in frame C, the user shifts his attention, diverting his focus away from the primary task to respond to ISTAR's interruption. The time between when ISTAR initiated the interruption and when it captured the user's attention is the interruption lag. ISTAR thanks the user for his response in frame D. Frame E depicts the completion of the interruption as ISTAR resumes its idling behavior. Frame E also shows the user resuming his original task. The resumption lag is computed from the completion of the interruption interaction to when the user resumes his original task.

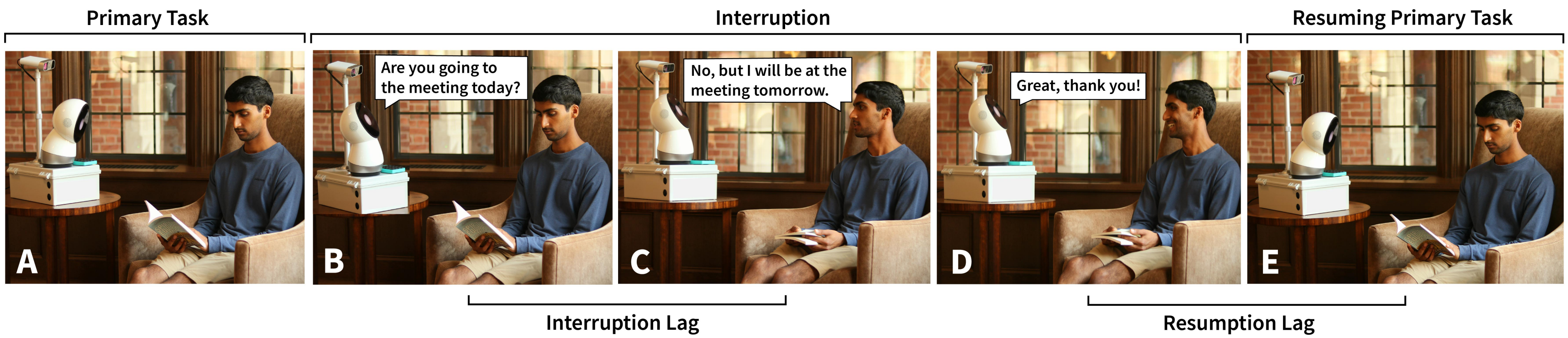


Figure 6.2: ISTAR Interruptions Sequence. A session involves the following: (A) the participant is occupied with a primary task while the robot is performing idling behavior; (B) the robot interrupts the user by asking them a work-related question; (C) the user responds to the robot's interruption; (D) the robot thanks the user for their response; finally, (E) the user resumes their original task. We define two primary metrics in Section 6.3 to measure resiliency to an interruption: interruption lag and resumption lag.

6.4 System

In the following sections, we describe the hardware and software components of our system prototype to achieve the aforementioned design goals and interaction.

6.4.1 Hardware

To achieve these interruption interactions, our system is comprised of six main hardware components as shown in Figure 6.3. We used the robot Jibo [168] which stands 11 inches tall and has 3 full-revolute axes designed for 360-degree movement. Jibo's hardware capabilities allowed us to program personified behaviors such as naturalistic gaze, pose, and movement. We included a compact PC that communicates with other hardware, monitors the overall system, and serves as the local data storage during our in-home system evaluation.

Survey evaluations by adults with ASD and employers (Section 6.5.1) suggest implementing interruptions that require a physical response. We included a numeric keypad to facilitate interruptions that prompt users to complete a mental task and enter their response into the keypad. Jibo and the keypad are fixed to the top of a plastic case containing the PC and all remaining hardware components that users do not interact with but support ISTAR's functionality. For the sensing required for in-home use, we mount an Azure Kinect [479] camera to a mast behind and two inches above Jibo's head to maximize the camera's field of vision. The Kinect also has a microphone array to capture audio during ISTAR training sessions.

We included several accessories to ensure the system is self-reliant in that it maintains power and internet connection once in the user's home. Each system is outfitted with a mobile router with a prepaid internet service plan for continuous WiFi connection. The router also enables automatic cloud-based data synchronization and remote control of the system for troubleshooting and system-monitoring purposes during our in-home evaluations. Additionally, the system is equipped with an uninterruptible battery power supply which serves as ISTAR's main charging station. This pack improves system robustness in the event of power outages.

With these components, ISTAR is a plug-and-play system that only requires connection to a power outlet in the user's home. Our hardware ensures self-reliance and self-containment. Considering rules for ergonomic and accessible design, we reduce the apparent complexity of the system by encasing its non-interfaceable components in the container which the robot and the external camera are mounted on.

6.4.2 Software

Interaction Components

We used a modular software architecture when creating the system to allow for individual components to be easily updated and improved. To achieve this modularity, we created the different components of our software as nodes in the Robot Operating System (ROS) [433].

The scheduling node determines when the system will capture an image. The images are captured on the system's Azure Kinect and then used as input to a pre-trained YOLO [480] neural network to predict the number of people in the system's field of view. If only one person is detected, then ISTAR delivers an interruption. If two or more people are detected, the system assumes that this is not a socially appropriate time to interrupt its user and it skips the planned interruption. Yet, the frequency of interruptions incrementally increases such that the number of interruptions within the designated time window remains the same. The time between interruptions is selected from a Gaussian distribution to prevent the user from predicting when the next interruption will occur.

When not delivering an interruption, Jibo silently looks at the floor. When it is prompted to deliver an interruption, Jibo looks up and plays a pre-recorded audio file of the interruption from its speakers. For interruptions that require a verbal response, Jibo waits for the user's verbal response which is then sent to the Google Speech-to-Text API so that the user's response can be transcribed. If the user does not respond within ten seconds, Jibo will reprompt them with the original question. After receiving their response, Jibo thanks the user and resumes silently looking at the floor.

Robustness for in-home study

Robots deployed in the home generally require significantly greater robustness than robots used in a lab setting. The unstructured environment of the home comes with many challenges, including the possibility of power outages, variable lighting conditions, and unexpected events that distract the user. To make our system robust to this unpredictable environment, we added software to inform us of the system's performance and the ability to remotely fix whatever problems may arise. This was achieved by using watchdog scripts and remote desktop applications.

The system has two watchdog scripts that run each day. The first script runs

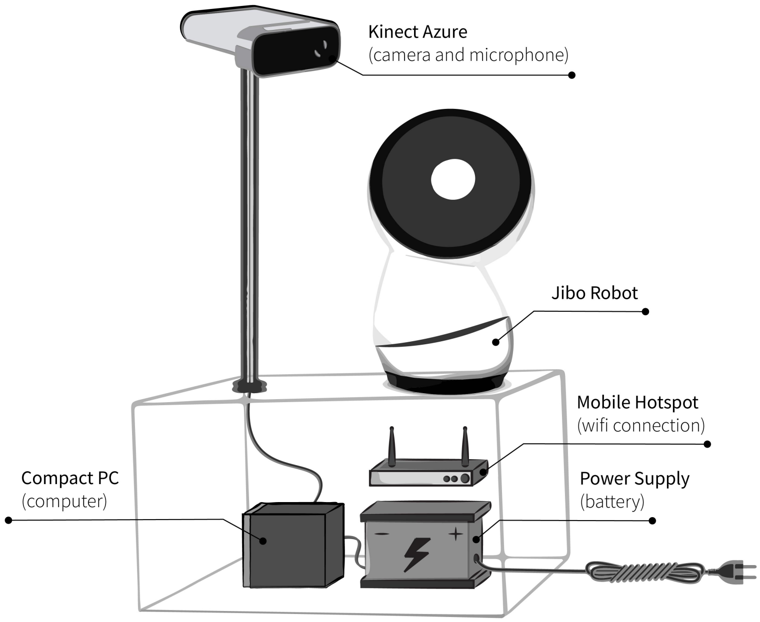


Figure 6.3: ISTAR Hardware. The system has a battery, compact computer, and mobile hotspot that are contained in a hard plastic case. An external camera and microphone are mounted on a mast above the robot's head. We later include a numeric keypad based on reports of common workplace interruptions experienced by adults with ASD in Section 6.5.

at the start of the training session. It verifies that the camera and microphone are successfully capturing images and sound, and that the PC is able to communicate with Jibo. The second watchdog script runs at the end of each day to check the size of the video, audio, and other files recorded to determine if the system turned off during the training session. It also checks the number of times that each type of interruption was delivered and the participant's responses to the interruptions. Each script notifies the research team detailing the success or failure of each of the components.

The system has two remote desktop applications [481, 482] installed to allow for remote configuration and debugging during the in-home evaluation. Remote access allows for remote configuration; the system can be delivered to the user's home and then configured completely without human contact.

6.5 Survey Evaluations of the Prototype

We conducted surveys of adults with ASD and employers to rapidly assess user acceptance of the system. We used the insights gathered from these assessments to improve the system before carrying out the more extensive in-home evaluations.

We showed three videos demonstrating ISTAR's operation. In the first video, ISTAR interrupted a user who was playing video games to ask if they would be willing to switch work shifts. In the second video, ISTAR interrupted a user watching a televised sports game to ask if they had completed a work report. In the third video, ISTAR interrupted a user washing dishes to ask where it would find an item in a grocery store.

Survey respondents were presented these demonstrations of ISTAR interrupting three different users. They evaluated the characteristics of the interruption, robot, and the overall interaction. Finally, participants were asked whether they would be willing to and how they would use the training system.

6.5.1 Results

We collected responses from 35 adults diagnosed with ASD and 13 employers of adults with ASD. A majority of the participating adults with ASD were students (89%). 31% were employed, and 26% were unemployed while 17% were actively looking for work. The remaining student-respondents were not seeking employment at the time surveyed.

Surveys of Adults with ASD

Of adults with ASD who were employed ($N = 11$), commonly reported workplace distractions included peer colleagues interrupting on matters unrelated to work (reported by 73%), supervisors interrupting on matters unrelated to work (55%), and environmental noise (e.g., a car honking its horn outside; 73%). When asked if ISTAR's interruptions were similar to those at the workplace, 23% responded that it was similar, 50% reported that it was somewhat similar, and 28% responded that it was different.

We explored how potential end-users would feel about having ISTAR in their home by asking if they would show this system when friends visited. 54% of all adults with ASD surveyed responded that they would show the system by interacting with it in front of their friends, 23% would just show the system to their friends, without demonstrating its functionality, 14% would ignore the system, and 9% would turn it off and place it out of sight when their friends came to visit. Using a 7-point Likert scale where 1 is *extremely uncomfortable* and 7 is *extremely comfortable*, participants reported they would be roughly equally comfortable interacting with ISTAR in their home ($M = 4.80, SD = 1.80$) as in their workplace ($M = 4.74, SD = 1.70$).

While 66% of adults with ASD did not find ISTAR and the interruptive interactions overstimulating, 13% found the robot's behavior and 8% found what the robot said overstimulating. For example, one respondent mentioned that it "speaks in a fast tone" and another said that "it kept moving and flashing." Several respondents mentioned that the robot distracted the user from their current task. One respondent said, "The information takes you out of and away from the current task." These observations from respondents align with our design goals because we want ISTAR to disengage the user from their task to practice responding to the interruptions.

In all, adults with ASD positively evaluated ISTAR's features and viewed the training interactions as valuable. 40% of participants reported that they would use this in-home system if it improved their prospects of getting a job, 34% of participants said that they would probably use it, and 14% said that they might or might not use it. The remaining 11% said that they would probably not use it.

Surveys of Employers

Generally, 80% of employers reported a difference in how adults with ASD handle workplace interruptions as compared to other workers. When asked to describe this difference, employers wrote that adults with ASD experience "difficulty concentrating or returning to [the primary] task" and that many "have adapted protocols on how to stay on or come back to [the primary] task."

Similar to adults with ASD, employers said they expect the most common workplace distractions to be environmental noise (reported by 77%) and peer co-workers on matters unrelated to work (reported by 69%). From employers' experiences, it took adults with ASD approximately 30 minutes and 40 seconds (SD = 38 minutes and 10 seconds) to return to their primary task, once interrupted. Adults with ASD and their employers differ in their perception of how long it takes for an adults with ASD to return to a primary task.

Limited significance should be placed on our survey respondents' estimation of the time it takes to resume tasks. Employers of adults with ASD reported that it took adults with ASD slightly over 30 minutes, on average, to return to their primary task, once interrupted. Whereas, employed adults with ASD reported taking about six minutes, on average, to resume their task. This five-fold difference could be due to employers making estimations based on observations of employee overall performance, recalling employees who took the longest to resume their tasks, and generalizations among employees.

6.5.2 Discussion

This evaluation establishes that ISTAR addresses a relevant and pressing problem, and could be accepted and utilized as a training platform. Most employers stated that employees with ASD handle workplace interruptions differently than other workers, and many adults with ASD said they would probably or definitely use our system if it would increase their prospects of getting a job. Survey results also indicate that most adults with ASD felt that they would be comfortable with ISTAR in their homes, even wanting to show off ISTAR and its interactions to a visiting friend. Most adults with ASD viewed ISTAR as friendly, approachable, and not overstimulating.

Insights collected from these surveys suggest that an interruptions training system should provide various types of interruptions to better resemble those frequently experienced in the workplace. We implemented three interruption modes on ISTAR, each of which demands a different skill or form of response. A *social* interruption requires a verbal and behavioral response to a user-directed question (e.g., turning to face the system, maintaining eye contact, and answering a robot-initiated question completely and appropriately). In contrast, a *task* interruption requires the user to physically interact with the system by typing in their response into a keypad. Last, an *environmental* interruption is a sound played through the robot's speakers and the expected behavior is for the user to ignore the entire interruption and continue their original task.

Each interaction with ISTAR begins in the same way: the robot transitions from its idle, sleep-like state and delivers an interruption prompt. The remainder of the interaction varies depending on the type of interruption. Given the wide variability in cognitive, communicative, and educational profiles among individuals with ASD, ISTAR does not assess the correctness of user responses. Instead, it focuses on creating consistent opportunities to practice regulation and task recovery, regardless of response content or form.

For **social interruptions**, the robot asks the user a question and then waits for them to verbally respond. If the user does not respond, it re-prompts the user with the same question. The robot then thanks the user when it receives a response. Examples of social interruptions include, "How do I get to the nearest train station?" and "In which aisle can I find pickles?"

To support task interruptions, we include a numeric keypad as illustrated in Figure 6.3. For **task interruptions**, the robot asks the user a question and requires them to type their numerical response into its keypad. ISTAR will re-prompt the user if they do not respond. Once the user types their response into the keypad, the robot thanks them for responding. Examples of task interruptions include, “Please enter in your zip code.” or “How many days are there until the weekend?”

For **environmental interruptions**, the robot plays a sound that one might typically find in a workplace environment, like the sound of a car driving by or cafeteria chatter. After the interrupting sound is finished, the robot then returns to its idling behavior of silently looking at the floor.

We designed a system capable of delivering robot-initiated interruptions and validated our design decisions through surveys with potential end-users. Most respondents with ASD indicated they would be willing to use ISTAR in their homes. An in-home setting is both practical and ethically appropriate, as it grants users greater control over their training environment. For example, users can choose when the robot is permitted to record or initiate training sessions, select which room to conduct the training in, and adjust system features to suit their preferences and environmental needs.

We improved the robot-initiated interruptions to better resemble interruptions commonly encountered in the workplace by implementing several types of interruptions. Based on the frequencies of workplace interruptions reported in our surveys, we configured the final system to interrupt users an average of 8 to 15 times in each two-hour daily training session. We also improved the content of the interruptions to make for more realistic and generalizable interactions.

6.6 In-Home Deployments and Evaluation

The best evaluation of this system is in the homes of adults with ASD. However, experiencing long-lasting improvement or behavioral change as a result of training would take several weeks to achieve [483]. Before we can fully evaluate the efficacy of ISTAR, we investigate whether adults with ASD will accept the system in their homes and continue to interact with its training prompts throughout a week-long study. The results of this evaluation can support longer-term deployments of ISTAR to explore lasting behavioral improvement in users.

After we received Institutional Review Board approval for the study, adults with ASD consented to participate by signing up through a website promoted via locally posted flyers. Due to the ongoing COVID-19 pandemic, special attention was given to ensuring that the system could be installed and operated independently by participants. Each ISTAR unit was delivered directly to participants’ homes and set up entirely by the users, without any in-person contact with the research team. Upon receiving the system, participants were encouraged to place ISTAR in a room where they spent most of their time and felt comfortable engaging in typical daily activities. They also specified the time windows during which ISTAR was permitted to initiate training sessions. During these sessions, ISTAR would "wake up" several times during the specified window to engage the participant in brief conversational interruptions. Each study concluded after ISTAR had been active in the participant's home for seven consecutive days.

6.6.1 Data Collection

Video and audio data recordings for all training interactions fully captured each ISTAR-given interruption the participant experienced, participant responses to the interruption, and their activities before and after responding to the interruption.

We performed four sets of annotations on each interruption given by ISTAR. Three researchers used ELAN [484] to timestamp when participants first turned their gaze away from their primary task after an interruption is given, then looked at the robot, turned their gaze away from the robot, and finally looked back at their primary task. At the beginning of this process, the transcriptions were evaluated twice for procedural errors. After the process completed, the inter-coder reliability was computed for 25% of all interruptions, randomly selected across participants and annotated by three coders. We evaluated the agreement between annotators because of the inherent ambiguity in assessing participant behavior in the noisy, unstructured home environment. The intraclass correlation coefficient was 0.95 and 0.90 for the time it takes the participant to look at the robot after an interruption is delivered (i.e., interruption lag) and the time it takes to look back at the primary task after an interruption is addressed (i.e., resumption lag), respectively.

To supplement these annotations, one member of the research staff transcribed objective characteristics of the participants' interactions using a survey. These transcriptions assessed the length of the participants' verbal responses to ISTAR, whether the participant resumed their original task or transitioned to a different task after an interruption, and how socially or physically demanding their tasks were before and after an interruption. These transcriptions were made using a series of objective binary questions, so computing agreement and multiple annotators were not necessary.

6.6.2 Participant Information

Twelve adults with ASD enrolled in this study. Two participants withdrew because of unrelated personal circumstances due to the pandemic. 8 males and 2 females, ranging from ages 20 to 42 ($M = 26.3$, $SD = 6.9$) years, completed this evaluation of ISTAR. Participants completed surveys to determine their level of education, employment status, AQ-10 score, and expectations of training with ISTAR using the Flow in Work Scale (FWS) [485]. Among the ten individuals who completed the study, nine participants completed the online survey and one participant required support from a caregiver to navigate the survey website and submit his responses.

Two participants were employed at the time of their study, five were unemployed and actively looking for employment, and three were not looking for employment. All participants had at least a secondary school experience with 80% having attended college or vocational training. Participants were high-functioning adults with a confirmed diagnosis of ASD and an average AQ-10 score of 4.6 ($SD = 1.6$). On a 5-point Likert scale, where 1 is *not easily at all* and 5 is *extremely easily*, participants reported being *somewhat easily* distracted ($M = 3.1$, $SD = 1.17$) from everyday interruptions. Responses to the FWS suggested that participants generally anticipated a moderate likelihood of success, found interacting with ISTAR to be interesting but not overwhelming, and viewed the training as a meaningful challenge they were eager to undertake. Specifically, participants reported moderate levels of fluency of performance ($M = 24.0$, $SD = 5.89$), absorption in activity ($M = 15.0$, $SD = 4.38$), and perceived fit between task demands and personal skills ($M = 13.0$, $SD = 4.05$).

6.6.3 Results

ISTAR delivered 841 interruptions in total. 12% of interruptions were removed from analysis because participants were not in the room to experience them. Each participant experienced an average of 73.2 total interruptions, 12.9 ($SD = 3.4$) interruptions per training session. In a workplace setting, we would define successfully handling environmental interruptions as seamlessly performing one's task despite the interruption. For social interruptions, we expect an employee to pause their task, maintain eye contact with the interrupter, and address the interrupter's question completely before resuming their task. For task interruptions, a verbal response is not necessary, but a complete and relevant response is. We also expect to observe reduced interruption and resumption lags throughout the training. This would indicate that users improve at switching between tasks and interruptions, and that ISTAR would be an effective system for achieving this improvement.

Handling Interruptions

According to these criteria, participants responded appropriately to 40% of all environmental interruptions experienced ($N = 237$), 98% of social interruptions, ($N = 250$) and 99% of task interruptions ($N = 245$). Across all social and task interruptions, participants had a high response rate to the interruptions, responding socially to 99% of the interruptions by sustaining eye contact, pausing their original task to attend to the interruption, or speaking to the robot. Interestingly, participants showed similar social behaviors for 60% of all environmental interruptions experienced.

A multiple linear regression calculated to predict interruption lag revealed a significant effect of the interruption type ($\beta = 2.37$, $p \leq 0.001$), AQ-10 score ($\beta = 0.45$, $p \leq 0.001$), and number of interruptions experienced into training with the system ($\beta = -0.01$, $p = 0.01$). The significant decrease in interruption lag as users continued to train with ISTAR shows that they attended more quickly to interruptions over time. A regression to predict resumption lag revealed a significant effect of interruption type ($\beta = -11.1$, $p \leq 0.001$) and AQ-10 score ($\beta = -1.02$, $p \leq 0.001$). Estimated coefficients are denoted as β.

Interruption and resumption lags were computed to compare the disruption caused by each type of interruption as measured in seconds (s). Participants' interruption lags were significantly shorter for environmental interruptions ($M = 2.24s$, $SD = 4.02s$) than for social interruptions ($M = 3.18s$, $SD = 3.45s$, $t = 2.66$, $p \leq 0.01$) and task interruptions ($M = 4.66s$, $SD = 4.44s$, $t = 6.00$, $p <= 0.001$). The interruption lags for social interruptions were also significantly shorter than for task interruptions ($t = -4.03$, $p \leq 0.001$).

Participants' resumption lags were significantly longer for environmental interruptions ($M = 15.86s$, $SD = 13.10s$) than for social interruptions ($M = 4.57s$, $SD = 6.82s$, $t = -11.57$, $p \leq 0.001$) and task interruptions ($M = 7.47s$, $SD = 6.89s$, $t = -8.35$, $p \leq 0.001$). The resumption lags for the task interruptions were also significantly longer than for the social interruption ($t = -4.49$, $p \leq 0.001$).

Perception of the System

At the end of their study, participants gave feedback on their experience with ISTAR by completing an online survey and interview. Using the Robotic Social Attributes Scale (RoSAS) [188], participants perceived ISTAR as warm, competent, and not discomforting to use. The terms popularly used to describe ISTAR were *social*, *responsive*, *interactive*, *capable*, and *organic*.

Participants additionally evaluated ISTAR as a training system. They reported on a 5-point Likert scale, where 1 is *none at all* and 5 is *a great deal*, that training with ISTAR improved their tolerance for interruptions experienced outside of their training sessions ($M = 3.3$, $SD = 1.3$). In interviews, two participants reported that training with ISTAR was valuable as they continued to look for employment in that "[ISTAR] would remind me of what I'd have to do in anticipation of interruptions, like prioritize [certain tasks]" or "remember what I was focused on before," and "[ISTAR] could help me with situations at work when I'm dealing mainly with frustration, like when handling multiple customers." One participant reported that ISTAR had already helped them in their current job: "Whenever I finished with [training with ISTAR], there have been times where there were interruptions [on the job] where I've gotten right back to work." For another participant, training with ISTAR made him reflect on the interruptions he gave to others: "So I'm a big interrupter. I interrupt in conversations, and it made me think about what I'm doing to others." Finally, on a 5-point Likert scale where 1 is *not relevant* and 5 is *extremely relevant*, participants reported that the training provided by ISTAR was relevant to handling real-world interruptions ($M = 3.9$, $SD = 0.93$).

6.6.4 Discussion

Following our results, we evaluated the success of ISTAR's design according to our design goals. ISTAR is **embodied** as a social robot to engage users in a greater capacity than would virtual technology or cellphone applications. Our evaluations suggest that ISTAR's physically co-present interruptions and socially-situated practice are likely to generalize to real-world interruptions. Embodiment allows for ISTAR's naturalistic gaze patterns and body movement that encouraged participants to practice their social responses. A caregiver remarked on the impact the physical presence of the system had on her daughter with ASD: "She absolutely loved it! As soon as [ISTAR] came into her apartment, it sparked her. She liked the way [ISTAR] moved, its personality, and she just came to life!"

As an **in-home** system, we emphasized the importance for individuals with ASD to intuitively and comfortably interact with ISTAR. By minimizing the design and interfaceable components of ISTAR's hardware, we gave users greater autonomy over where, how, and when they interacted with the system. Our results confirmed that users would be comfortable interacting with ISTAR in their homes, even to the extent that they would show off ISTAR and its interactions to a visiting friend. Most participants believed that ISTAR was friendly, approachable, and not overstimulating.

In addition, ISTAR operated **autonomously** for a total of 1680 hours, successfully delivering 70 training sessions. Autonomous interactions present substantial challenges in computational perception and system control to create meaningful social-skills interventions. Yet, our implementation of watchdog scripts and remote software allowed us to ensure participant data is properly collected and stored during the in-home evaluation. Furthermore, due to the COVID-19 pandemic, the system was designed to be intuitive to install and use. All systems were deployed and setup completely without the research team making direct contact with users or their homes.

Interactions with ISTAR are **realistic**. In designing a system to improve employability through interruptions training, it is intuitive to have only job-specific content. However, not all interruptions are familiar to most jobs or individuals that are not yet employed, and would be aligned with reports of the most distracting interruptions in Section 6.5. Adults with ASD and employers evaluated the interruptions of an ISTAR prototype that produced only work-related interruptions as being only "somewhat similar" to real workplace interruptions. As a result, we vary the physical and social demands of interruptions relevant to most workplaces by implementing three types of interruptions: social, task, and environmental. All employed adults with ASD that participated in the in-home evaluations of the final system reported instances in which they felt they handled real workplace interruptions better due to the interruptions training they experienced with ISTAR. As this work is an early step towards understanding the potential for an in-home social robot for adults with ASD, a longer-term study with a larger sample is needed to investigate whether ISTAR will generalize to workplaces or human-human interactions.

We did not expect significant behavioral change in a week-long study to indicate efficacy of our system. Surprisingly, our in-home evaluation demonstrated that training with ISTAR significantly improved participants' ability to attend more quickly to interruptions over time. Based on computed lags, ISTAR's various types of interruptions produced significantly different disruptions and participant responses. Still, participants practiced appropriate social behaviors to almost every interruption experienced throughout their entire study such as sustaining eye contact, pausing and returning to their original activities, and speaking with the robot. In all, interactions with ISTAR are productive and can be an effective system for improving interruptions tolerance in adults with ASD.

6.7 Summary

This chapter introduced a social robot designed to help adults with ASD practice handling workplace-relevant interruptions. Grounded in the real-world challenges faced by individuals with ASD—particularly around sustaining employment and coping with unpredictability—ISTAR aims to provide socially meaningful, in-situ training through brief, interruptive interactions. The system was designed with end-user input, validated through survey feedback, and deployed in a fully contactless manner due to the COVID-19 pandemic. Our evaluations show that participants readily accepted ISTAR into their homes, found the training relevant and valuable, and showed significant improvements in their ability to manage interruptions. By enhancing users' resilience to disruptions, ISTAR contributes to the broader goal of supporting social regulation and employment readiness within an underserved population.

Within the broader goals of this dissertation, the work presented in this chapter builds on lessons learned from developing robots during the global pandemic to create a system that can be delivered entirely contact-free and set up by users without the need for technical expertise. It represents the first in-home, robotic intervention designed specifically for adults with ASD. It further demonstrates how a robot can provide meaningful social exposure to support the practice of a desirable regulation skill over time.

Yet, ISTAR features rote, pre-scripted practice. Although effective for ensuring controlled training and broad usability, this approach only partially addresses the broader aims of this dissertation. It omits three essential characteristics of real-world social interaction: (1) reciprocal dynamics, (2) contextual timing, and (3) organic social feedback. Our work following the ISTAR deployment examines each of these components in greater depth. For example, we explored how robots can determine the appropriate timing to engage users [33,34]—an especially important consideration for systems embedded in users' daily routines at home.[2] The following chapter explores how robots might move beyond one-directional prompts to participate in back-and-forth exchanges that more closely mirror natural conversation. We also examine how robots can provide directive feedback—either on their own social behavior (Chapter 7) or on users' social progress to deliver more explicitly instructive interventions (Chapter 8). Unlike cognitive domains (e.g., teaching math skills), where correctness is objectively defined, social behavior resists straightforward evaluation, is highly

[2]For clarity, these studies are not presented as dedicated chapters in this dissertation. However, its proposed models are incorporated into subsequent deployments and referenced where applied.

context-dependent, and is often intertwined with personal identity—making feedback more complex and delicate to deliver effectively. Together, the following two chapters (7 and 8) present our developments toward a second robot-assisted intervention for adults with ASD, focused on supporting a novel set of social regulation skills.

Chapter 7

A Grounded Observer Framework for Establishing Guardrails for Foundation Models in Socially Sensitive Domains

As noted in Chapter 2, systems that offer novelty, personalization, and adaptability are more likely to sustain engagement over extended periods than systems that rely on scripted or rule-based interactions. With the advent of foundation models (large-scale pre-trained models capable of generating diverse, contextually appropriate responses; [486]), we can now envision systems that more feasibly support long-term, dynamic engagement with users.

However, as foundation models increasingly permeate sensitive domains such as healthcare, finance, and mental health, ensuring that their behavior aligns with desired outcomes, ethical norms, and social expectations becomes not just important, but *required*. Yet, due to the opaque, high-dimensional nature of these models, conventional methods for constraining agent behavior (typically designed for discrete, low-dimensional state and action spaces) are ill-suited for governing the dynamic and generative outputs of foundation models.

To address this gap, this chapter introduces the grounded observer framework, a novel method for shaping the behavior of foundation models through real-time evaluation and constraint. Drawing inspiration from action-selection mechanisms in robotics and human-in-the-loop control paradigms, this approach shifts the focus from altering the model's internal mechanisms to regulating its observable behavior. By continuously assessing low-level behavioral features, the observer can enforce symbolic constraints that reflect both contextual demands and user-defined preferences. This allows for real-time variability while still offering behavioral consistency and oversight.

We present small talk as both a valuable testbed and an emerging frontier for development, one that reveals challenges in applying behavioral guardrails to foundation models. To demonstrate our proposed framework, we develop a system that sustains casual, socially appropriate conversation (i.e., small talk) and integrate it on a physically embodied robot. The robot engages in novel, unscripted interactions with human users, adapting its conversational style and behaviors in real time based on observer feedback. This implementation explores how foundation models can support socially responsive and adaptive robot behavior for sustained interactions in complex, real-world settings. The grounded observer framework further ensures that such generative systems are deployed in a more responsible, ethical, and socially appropriate manner.

7.1 Introduction

Foundation models are rapidly being integrated into various fields, from medical diagnostics and financial predictions to socially sensitive areas such as education, mental healthcare, and support for individuals with disabilities. Despite being aware of the inherent risks of AI hallucinations, misinformation, and bias, a recent large-scale global study revealed that 66% of respondents are still willing to use this nascent technology in sensitive areas such as personal advice and relationship counseling [487]. This paradox highlights the immense potential benefits of these models in addressing societal challenges while also underscoring the current concerns. A significant issue tempers the widespread adoption of these tools: the lack of comprehensive guardrails to prevent undesired behavior and ensure reliable outcomes.

In fields where accuracy and reliability are paramount, such as healthcare and finance, the consequences of errors can be severe. Yet, in socially sensitive domains, where the parameters of success are less tangible, the impact of missteps can be as profound. For example, a system intended to provide calming techniques in a clinic waiting room could exacerbate anxiety if it delivers generic or poorly timed suggestions. If it fails to recognize the urgency or context of a patient's distress, it may offer advice that feels dismissive or irrelevant, potentially increasing the patient's anxiety. In light of such effects, foundation models should have robust guardrails to protect users and the system's integrity.

Designing usable systems that impose limits on foundation models involves two key challenges. First, foundation models are based on statistical learning from vast datasets, making their internal mechanisms complex and opaque. Traditional rule-based systems use symbolic representations, which are formal and interpretable but not directly compatible with the statistical nature of foundation models. This difficulty is compounded when integrating symbolic rule-based systems that map human concepts into precise rules, a challenge akin to reconciling statistical learning mechanisms with symbolic representation systems. While neurosymbolic approaches that aim to blend statistical and symbolic methods are being explored (e.g., [488]), effective integration remains an open area of research.

Second, foundation models must be able to adapt their behavior in real-time to the unique needs and contexts of individual users [489, 490]. Static, predefined rules often do not address the dynamic and nuanced nature of personal interactions [491]. For instance, a large language model (LLM)[1] for mental health support must respond appropriately to a user's current emotional state and context. A static rule-based approach may fail to provide suitable support during a crisis or tailor interactions based on ongoing conversations, highlighting the need for real-time adaptability to meet individual user needs.

These two challenges are not unique to foundation models but manifest in other areas, such as robotics. In action selection for robot systems, an agent must decide on actions to take, often using large-scale statistical models, while adhering to user-specified rules, such as "don't touch the stove." Addressing this involves techniques known as shielding [492] and interactive policy shaping [493]. Shielding techniques prevent particular actions from being executed, effectively restricting the robot's behavior, while interactive policy shaping modifies the action selection policy in real time based on user input or situational changes. These approaches aim to reconcile the flexibility of statistical models with the necessity of adhering to predefined constraints [494], reflecting similar challenges faced in the context of foundation models.

Drawing inspiration from robotic action selection techniques, we propose a framework for constraining foundation model behavior that offers both behavioral guarantees and real-time variability. This method involves a grounded observer that continuously assesses the underlying model's candidate actions based on low-level behavioral characteristics, makes dynamic adjustments to the model's action generation, and provides feedback directives to ensure the behavior remains contextually appropriate and effective.

In this chapter, we present the conceptual framework of the grounded observer for establishing guardrails for foundation models. We apply this framework to build agents capable of small talk, a task that requires nuanced social sensitivity to ensure continued appropriateness and relevance. This case study of small talk demonstrates

[1] A LLM is a type of artificial intelligence system trained on massive datasets of text to generate, understand, and respond to human language. While all LLMs are foundation models, not all foundation models are limited to language.

how the grounded observer can impose precise constraints on LLM behavior in highly subjective contexts and challenge the typically informative and assistive nature of these models. We also demonstrate that this method leads to more positive and socially appropriate interactions when integrated into a robot where its embodiment amplifies social impacts. Lastly, beyond small talk, we explore how this technique can be applied to create guidelines in various socially sensitive domains.

7.2 Related Work

Given their complexity and the vast datasets they are trained on, ensuring that foundation models behave in predictable and socially acceptable ways is a significant challenge. Researchers have explored approaches to impose constraints on these models, each with strengths and limitations.

7.2.1 Prompt Engineering

The current standard for constraining model behavior is having a good prompt. While crafting specific input prompts has shown promise in many applications [495–497], it has significant limitations when it comes to robustly constraining agent behaviors, especially in complex, dynamic, and sensitive contexts.

Lack of Robustness. While specific prompts can guide the model in controlled scenarios, they often fail to generalize across different contexts and variations. A prompt that works well in one situation might produce unexpected or undesirable results in another, leading to inconsistent behavior [498, 499].

Context Sensitivity. Foundation models are highly sensitive to the context provided by prompts. Small changes in phrasing can lead to significantly different outputs, making it challenging to predict and control the model's behavior reliably [500, 501]. This sensitivity can be particularly problematic in dynamic environments where the context is continuously changing.

Inability to Enforce Hard Constraints. Prompt engineering cannot enforce hard constraints on model behavior. While prompts can suggest or guide the model toward certain behaviors, they cannot guarantee that it will always comply with these suggestions [502]. This limitation is critical in applications where strict adherence to ethical guidelines or safety protocols is necessary.

Translating to Real-World Behavior. Many real-world scenarios involve ambiguous and complex situations that are difficult to capture with prompts [503]. For instance, ensuring that an LLM provides appropriate mental health support requires understanding and responding to nuanced emotional cues, which cannot be fully encapsulated in a prompt. In such cases, prompt engineering alone cannot ensure reliable and sensitive behavior.

Temporal Constraints. Prompt engineering does not inherently support temporal constraints, where the desired behavior depends on the sequence and timing of interactions [504, 505]. For example, maintaining consistent behavior over multiple exchanges with a user is challenging to achieve through prompt design alone.

7.2.2 Constrained Reinforcement Learning

Constrained reinforcement learning (CRL) enhances traditional RL by integrating predefined constraints to ensure agents operate within specific safety, ethical, or operational boundaries. While traditional RL focuses solely on maximizing cumulative rewards, CRL incorporates additional constraints as hard limits (e.g., avoiding unsafe actions) or soft constraints (e.g., minimizing deviation from desired behaviors). CRL incorporates inductive biases through logical rules that govern the agent's behavior, applying these constraints directly to states and actions or modifying the reward function to align with the defined limits [506].

A notable approach within CRL is shielded RL, which employs user-defined policy overrides, or "shields," to restrict certain actions based on specific conditions, thereby minimally disrupting the RL model while enforcing desired behaviors [507]. However, shielded RL typically relies on a dynamic model and repairing existing policies rather than adapting to evolving preferences. In contexts such as personalized healthcare or companionship, a flexible approach to adapt policies to meet context-specific needs in real-time is more suitable.

7.2.3 Transparent Matrix Overlays

Transparent Matrix Overlays (TMOs) is a promising technique for real-time modification of agent behavior by integrating user directives as symbolic constraints on a robot's policy [508]. This approach merges concepts from CRL and shielded RL, leveraging symbolic reasoning to enhance flexibility in behavioral adaptation.

Demonstrated through a simulated collaborative cooking task [508], TMOs allowed adjustments to a robot's policy without requiring extensive retraining. By applying logical rules and user-specific directives as temporary constraints, TMOs facilitated immediate changes in behavior to align with evolving user preferences. This method contrasts with traditional CRL techniques, which often require substantial retraining to incorporate new constraints, and shielded RL methods that focus on policy repairs rather than accommodating real-time preference changes. This approach balances the stability of learned behaviors with the flexibility required to meet new and evolving preferences, making it a valuable tool for interactive systems.

One limitation of TMOs is the reliance on hand-crafted predicates and classifiers. In the current implementation, these elements are manually designed to define constraints and directives. While this method works within controlled environments, it constrains the flexibility of the TMO approach. The assumptions of having a relatively simple, discrete state space, deterministic actions, and non-parallel task completion further simplify the scenario. Real-world applications often involve more dynamic and complex environments where these assumptions may not hold.

7.2.4 State and Action Space Abstraction

Most action selection mechanisms, like TMOs, assume a known, discrete, or discretized state space with well-defined actions. However, for foundation models, an action selection mechanism must handle continuous and possibly infinite state spaces where iterating through all possible actions or states may be impractical [509]. This requires rules that can overlay abstracted state representations or symbolic predicates to approximate the agent's internal state and action space. Instead of exhaustively evaluating every action, the agent can use these overlays to focus on a manageable subset of candidate actions or employ probabilistic sampling techniques within the space emphasized by the overlays. Furthermore, creating overlays that are compatible with the diverse and proprietary architectures of various off-the-shelf foundation models can be challenging. Each model may have unique internal representations and decision-making processes, making it difficult to design universal overlays that function effectively. This variability requires overlays to be abstracted to a level that supersedes differences in how proprietary architectures handle context, manage memory, and generate responses [510].

7.3 The Grounded Observer Framework

Social behavior is inherently emergent and complex. However, in many cases, appropriate behavior can be guided by simple rules. Just as TMOs embed rules to control behavior, we can apply similar principles to ensure that foundation models exhibit appropriate social behavior. Foundation models are analogous to the action policies generated—they are statistical models that are expensive to generate, difficult to dissect, and opaque to inspection. By imposing transparent and adaptive constraints, we can manage and direct these models to align with desired outcomes in socially sensitive domains. This can be achieved by evaluating a model's output through context-based rules and providing feedback to guide the model toward more appropriate behaviors.

7.3.1 Overview of the Framework

We begin with a foundation model, referred to as the *base* model in Figure 7.1, which generates actions in response to environmental or user inputs. Depending on the type of model, these actions can take the form of text, images, or other outputs. To provide a clear overview in this section, we will focus on LLMs, assuming that both the model's inputs and outputs are in text form, though other modalities are also applicable. To evaluate the base model's actions, feature extractors convert these actions and the surrounding context into numerical features. These features can then be analyzed as scores based on the characteristics we want to evaluate. Depending on the scenario, these extractors may also incorporate inputs from high-level planners or context observers. For example, a feature extractor could be designed to quantify the politeness of the model's text output.

These contextual features are evaluated against IFTTT (If This, Then That) rules, which function as overlays on the model's actions. Think of these rules as semi-transparent sheets on an overhead projector: you can stack, prioritize, or remove them to adjust the view without altering the original image. Similarly, these rules can be adjusted without extensive changes to the base model.

High-level descriptors—summaries of how well proposed actions align with the overlays—are given by each overlay rule in a fixed text structure. These descriptors pinpoint areas where proposed actions comply with or deviate from the established rules. For instance, a rule about politeness might provide a directive like "tone is too polite," while a rule that assesses user frustration could direct the model to include more empathetic language. Each overlay also produces a score indicating the degree of deviation from the rule. These scores highlight more severe violations by using methods such as ranking or keyword tagging in the directives.

An *observer*, a separate model instance, receives these directives, then combines and translates them into actionable feedback for the base model. For example, if

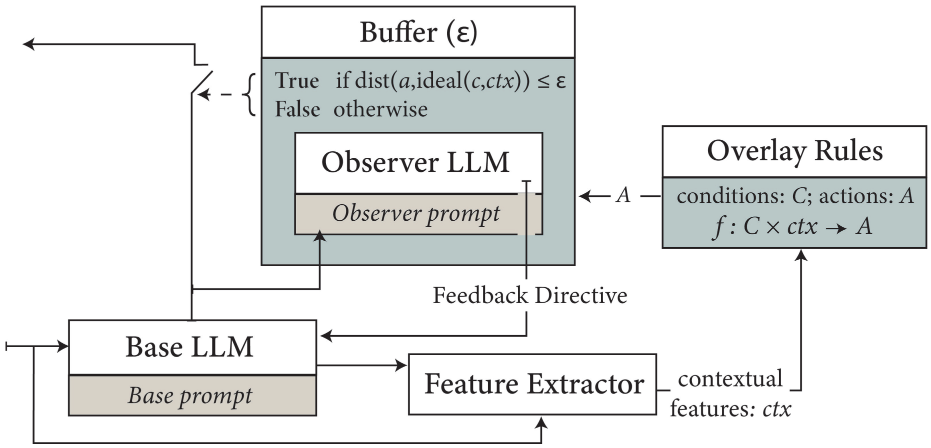


Figure 7.1: The grounded observer monitors a base model's behavior to ensure responses adhere to overlay constraints.

a directive indicates that the tone is too polite, the feedback might be, "The reply was overly formal. Please adopt a more casual tone." To our earlier analogy of the overlays as clear sheets on an overhead projector, the observer looks through these transparent layers to identify candidate areas of the action space while other areas are obscured or "blocked" from selection. Just as you can stack, prioritize, or remove sheets on a projector without changing the base image, these overlay rules can be easily adjusted without permanently altering the underlying model. Ultimately, the overlay rules map the extracted features to specific actions, delineating which actions the base model can or cannot take in a given context.

7.3.2 Action Filtering

A buffer acts as a gatekeeper, as shown in Figure 7.1, determining whether a proposed action should be accepted. Each overlay can be assigned a rigidity parameter (depicted as ϵ) that defines how strictly the model must adhere to the rule. Essentially, in reference to the overhead projector analogy, this parameter controls the translucency of an overlay. Instead of enforcing a strict binary compliance—where actions either fully meet the overlays or not—rigidity offers a gradient of compliance or a buffer around proposed actions.

For highly rigid overlays, compliance is strictly enforced. If an action or response deviates from the specified rules, the base model is required to regenerate new candidate actions. This ensures that only actions meeting the strict criteria are considered. For instance, if an overlay rule demands that responses must be empathetic, any response lacking empathy would lead to the base model generating alternative responses that conform to this requirement. In Section 7.3.4, we outline three ways in which overlay rules can "demand" compliance.

Overlays with lower rigidity act like more translucent filters, allowing actions that partially satisfy a constraint to pass through. Rather than strictly enforcing binary compliance, the observer can rank or prioritize these partially aligned responses and accept them within a permissible margin. For instance, if the overlay demands empathetic responses, the model may still accept a reply that conveys moderate empathy if it is otherwise contextually appropriate.

This flexibility helps manage the model's load and processing time when correcting its actions. For non-critical constraints, lower rigidity reduces overhead by avoiding additional correction cycles. In contrast, high-rigidity overlays are reserved for critical conditions, where strict compliance is essential. In both cases, a buffer can cap the number of action regeneration attempts to prevent excessive resource consumption while enforcing the necessary constraints.

7.3.3 Feedback Directives

The observer utilizes the overlay descriptors and rigidity to create targeted feedback prompts to the base model. We incorporate two types of feedback:

Implicit feedback notes that the action is acceptable but offers constructive advice for improving subsequent actions. For example, if the actions are near compliance but not perfect, implicit feedback may recommend minor adjustments, such as modifying tone or phrasing. Suppose the base model generates a response that is mostly empathetic but could be softer in tone. The implicit feedback might suggest: "Consider using a gentler tone in your responses." This allows the base model to refine its output in future iterations.

Forced feedback is employed when the base model's actions significantly deviate from the overlay constraints. When the descriptors reveal substantial misalignment with the overlay rules, the observer generates a more directive prompt, instructing the base model to focus on specific improvements until it fully complies with the constraints. The observer may issue several rounds of feedback if needed until proposed actions meet the overlay requirements.

Overall, this feedback loop ensures that the base model continually aligns with the overlays by translating its performance on specific rules into clear instructions. In the next section, we apply this framework and demonstrate the role of each component within a social context.

7.3.4 Examples of Overlay Types

Overlays can be defined in various ways. Here, we present three distinct approaches, each yielding a different class of rules: prohibitory, transfer, and permissive. In this section, *model* refers to the base model, while *agent* denotes the system (robotic or otherwise) into which the model is integrated.

Prohibitory overlays are designed to restrict or minimize the selection of certain actions in specific states. Their primary function is to enforce constraints that discourage the model from choosing these actions. When strong constraints are required, prohibitory overlays can assign a zero probability to specific actions, effectively removing them from the decision space. This hard exclusion ensures that such actions can never be selected, even when no better alternatives are available, thus guaranteeing compliance with critical safety or ethical constraints. In less critical scenarios, prohibitory overlays may assign very low weights to undesirable actions rather than removing them entirely. This soft discouragement significantly reduces the chance of those actions being selected while keeping them available as fallback options. For example, in a navigation task, the robot might prefer routes that avoid passing through crowded areas. A prohibitory overlay could assign lower weights to actions that enter high-traffic zones, making them unlikely to be chosen under normal circumstances. However, if no clear alternative exists, such as during an emergency or detour,those paths remain technically available.

Transfer overlays, on the other hand, are used to shift the model's preference from one action to another based on specific conditions. This overlay type reallocates preference by assigning a low value to a source action and a high value to a target action. The effect is a clear preference shift, making the target action much more likely to be chosen while the source action is less favored. For example, if an agent needs to gradually move from preferring action A to action B, a transfer overlay would facilitate this change by decreasing the weight of action A and increasing the weight of B.

Permissive overlays are intended to encourage the selection of certain actions under specific conditions. Unlike prohibitory overlays that aim to reduce action probabilities, permissive overlays increase the likelihood of desirable actions by assigning them a high value if they meet the overlay's conditions. Actions that do not meet these conditions receive a standard value, meaning they are neither particularly encouraged nor discouraged. This type of overlay is used to promote actions that align with specific goals or strategies. For example, if a model is supposed to prioritize actions that lead to a positive outcome or strategic advantage, a permissive overlay would assign high values to these actions, thereby increasing their chances of being selected.

In summary, prohibitory overlays are used to discourage certain actions, transfer overlays to shift preferences between actions, and permissive overlays to encourage actions that align with desired conditions or goals.

7.4 Technical Demonstration: Developing Agents Capable of Small Talk

Imagine a modern care home for the elderly where a state-of-the-art robotic assistant, designed to enhance residents' well-being, manages routine healthcare tasks. Alex, a resident, seeks a connection beyond the daily routine and attempts to chat with the robot:

ALEX: Hi CareBot, how's it going?
BOT: Hello. How may I help you?
ALEX: Oh, just making conversation. Anything interesting happen in your world?
BOT: I have access to a vast database of news articles. Would you like information on a specific topic?
ALEX: No, never mind that. The weather will be nice this weekend. How would you spend it?
BOT: The weather forecast for the upcoming weekend predicts clear skies and mild temperatures, making it an ideal time for outdoor activities. Expect daytime highs around 75°F, 24°C and comfortable evening lows of 60°F, 15°C...

An essential aspect of designing intelligent systems today is to imbue some level of conversational behavior [511], particularly for socially interactive systems in personal, long-term settings [503]. Despite the potential for these agents to elicit meaningful interactions, the dialogue above exemplifies a common shortcoming: Alex seeks casual conversation with the robot, but it instead redirects the conversation toward programmed functionalities, insisting information and task-oriented assistance. Failure to meet social needs when deployed to autonomously operate with vulnerable populations, as seen in autism therapy or eldercare, can pose safety-critical concerns [512]. These limitations may impact mental health, hinder therapeutic progress, or create risks in sensitive care environments.

Small talk transcends the conventional definition of conversation. Unlike the functional aspects of conveying information or assistance, small talk acts as a social lubricant by fostering rapport and trust [513]. As a specific form of conversation, it has distinct traits [514, 515]: *brevity*, concise responses avoiding elaboration; *tone*, maintaining light-hearted, informal interactions; *non-specificity*, focusing on broad topics rather than details; and *thematic coherence*, ensuring relevance and continuity. Nonetheless, there is no strict formula as small talk is inherently flexible and context-dependent. Its fluid nature presents a significant challenge for LLMs, which typically rely on structured and well-defined question-answer patterns.

A skilled conversationalist not only learns their partner's preferences over time but also adapts to them in real-time, using naturalistic cues that may be linguistic, implicit, and contextual. For intelligent agents, this means they must swiftly adjust their policies in response to high-level, imprecise, or evolving directives conveyed through natural language. Therefore, we present a proof-of-concept, technical demonstration of how a grounded observer can dynamically shape an agent's behavior while adhering to high-level directives in the highly subjective social context of small talk.

7.4.1 Current Challenges in LLM Small Talk

We conducted an initial study to determine the extent to which small talk poses a challenge for LLMs.[2] Three volunteers engaged in 50 conversations each with three distinct state-of-the-art LLMs. Each model had the initial system prompt describing the role as a "friendly companion who engages in casual, small talk," with the prior listed criteria definitions. The selected LLMs are GPT-3.5 [516], for its large-scale language generation capabilities, Gemini Pro [517], for its context-aware bidirectional approach, and LLaMA-2 [518], an autoregressive transformer model fine-tuned on prompt-response pairs.

Data Collection

The order in which the participants interacted with the LLMs was randomized to mitigate potential order effects. Additionally, conversations lasted at least ten turns, and the interactions occurred over 15 days to allow for conversational variability. The participants engaged with each LLM through a command line interface, unaware of the LLM's name to prevent bias from prior knowledge or familiarity. Following each

[2] All methods described in this chapter were preregistered and received IRB approval.

conversation, assistants rated the ease of each conversation and provided open-ended feedback.

Two research assistants annotated the dataset. These raters were blind to the response speaker and evaluated responses based on recognized small talk criteria: brevity, tone, specificity, and coherence. Evaluations for each response based on these criteria were provided on a 5-point Likert scale, ranging from (1) *very concise* to (5) *very wordy*, *very negative* to *very positive*, *very general* to *very specific*, and *definitely not coherent* to *definitely coherent* [519].

Interlocutors typically have multiple goals in conversation [520, 521]. Even in casual small talk, where there are no task-oriented goals, interlocutors have various conversational goals such as conveying emotion and continuing the conversation [522]. Therefore, each LLM-generated response was further categorized by its conversational motives: informative, assistive, expressive, or person-directed. Definitions and examples for each of these categories were provided to the annotators [519]. As a single response can intersect with more than one category, the annotators rated the response for each motive using a 5-point Likert scale.

- *Informative*: Responses provide factual information, answer queries, or offer guidance related to specific tasks. For example, "I disagree. The forecast says it will be stormy this weekend."
- *Assistive*: Assistance-based responses provide help, guidance, or support to the user. For instance, "I'm sorry to hear that your car has broken down. How can I help?"
- *Expressive*: Expressive responses convey emotions, sentiments, or personal opinions. For example, "I recently visited a beautiful mountain resort. The scenery was breathtaking, especially during sunrise."
- *Person-directed*: These responses stimulate further discussion, invite the other person to share more, or ask questions to continue the conversation. For example, "What will you do with your time off from work?"

We acknowledge that these do not encompass the full spectrum of potential motives in dialogue. Rather, they were selected to provide a structured framework for analysis and interpretation of the suitability of LLMs for casual small talk.

Importantly, all participants were not familiar with the objectives of the present research to ensure unbiased assessments. This study protocol and hypotheses were preregistered [519] and received university clearance. Formal instructions and definitions presented to the participants were published on the Open Science Framework before data collection [523].

Results

A total of 150 conversations were transcribed, yielding an average of 10.31 responses per conversation ($SD = 1.13$). This led to a total of 1547 annotated responses.

We calculated the inter-rater reliability for a randomly selected subset of 20 conversations, constituting 13.3% of the total dataset. This assessment was deemed necessary due to the inherent ambiguity in evaluating the subjective qualities of responses. Inter-rater reliability was calculated using contingency tables, employing Cohen's Kappa (κ), with the observed agreement and the distribution of ratings for each rater. The resulting κ values were 0.81 for brevity, 0.78 for tone, 0.74 for specificity, and 0.65 for coherence.

A response may have multiple motives. Thus, we normalized ratings within the four conversational motives to a scale between 0 and 1. Then, we assessed agreement between raters using the intraclass correlation coefficient (ICC). The computed ICC values were 0.89, 0.86, 0.91, and 0.93 for the informative, assistive, expressive, and person-directed motives, respectively, indicating good to excellent agreement.

Human vs. LLM Comparison. We utilized paired dependent t-tests to assess the differences between the LLMs' and humans' responses across the four small talk criteria and four conversational motives. A conventional significance level of 0.05 was employed, and resulting p-values were Holm-corrected to control the familywise error rate.

The results revealed a significant difference in brevity ($t = 86.78$, $p \leq 0.0001$) between the LLM responses ($M = 4.55$, $SD = 0.97$) and human responses ($M = 1.23$, $SD = 0.54$), tone ($t = 1.70$, $p = 0.04$) between the LLM ($M = 3.02$, $SD = 0.33$) and human responses ($M = 2.99$, $SD = 0.52$), specificity ($t = 58.06$, $p \leq 0.0001$) between the LLM ($M = 4.54$, $SD = 1.09$) and human responses ($M = 1.66$, $SD = 1.02$), and thematic coherence ($t = -55.72$, $p \leq 0.0001$) between the LLM ($M = 1.88$, $SD = 1.23$) and human responses ($M = 4.56$, $SD = 0.89$). Together, this suggests that LLM-generated responses were considerably less concise, slightly more positive, more specific, and less thematically coherent than human responses.

We further observed statistically significant differences among all four conversa-

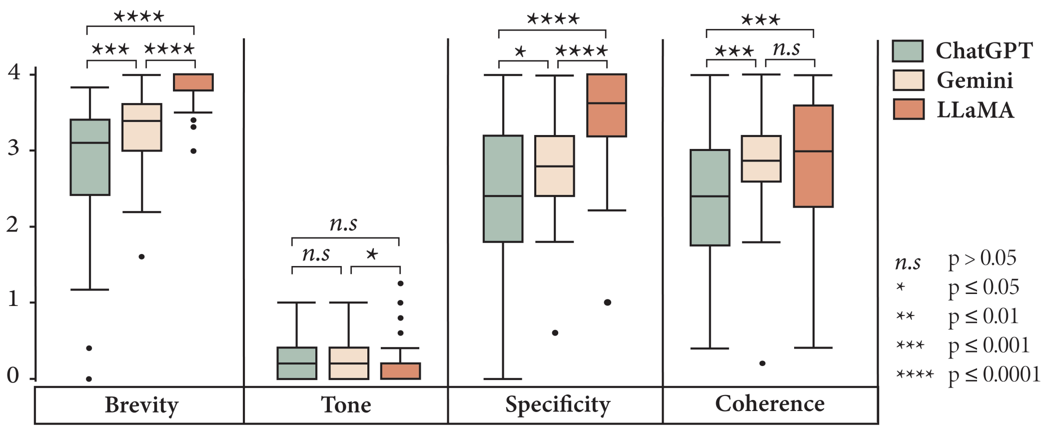


Figure 7.2: Human-Likeness of LLMs. This graph illustrates the extent of human likeness displayed by three LLMs, scored from 0 (no difference between human and model responses) to 4 (highest absolute difference). Each score reflects the similarity of the model's small talk to that of the participants.

tional motives: informative ($t = 25.67$, $p \leq 0.0001$) between the LLM ($M = 0.37$, $SD = 0.39$) and human ($M = 0.01$, $SD = 0.06$), assistive ($t = 24.51$, $p \leq 0.0001$) between the LLM responses ($M = 0.31$, $SD = 0.35$) and human responses ($M = 0.00$, $SD = 0.00$), expressive ($t = -24.22$, $p \leq 0.0001$) between the LLM ($M = 0.16$, $SD = 0.24$) and human ($M = 0.60$, $SD = 0.45$), and person-directed ($t = -12.815$, $p \leq 0.0001$) between the LLM responses ($M = 0.16$, $SD = 0.27$) and human responses ($M = 0.40$, $SD = 0.45$). In all, this suggests that LLM-generated responses were significantly more informative and assistive, and less expressive and person-directed as compared to human responses.

Comparison Between LLMs. We assessed the behavior of the three LLMs during the small talk interactions by comparing each pair of LLMs using the Wilcoxon method and Holm-corrected significances. Among the four criteria, ChatGPT 3.5 generated responses that were more consistent with our definition of small talk in that its responses were significantly more concise than LLaMA ($Z = 12.74$, $p \leq 0.0001$) and Gemini Pro ($Z = -8.81$, $p \leq 0.0001$), less specific than LLaMA ($Z = -10.21$, $p \leq 0.0001$) and Gemini Pro ($Z = -6.79$, $p \leq 0.0001$), and more thematically coherent than LLaMA ($Z = 5.51$, $p \leq 0.0001$) and Gemini Pro ($Z = 12.37$, $p \leq 0.0001$).

We determined the degree of similarity between LLM behavior and human responses by computing the absolute difference in their average scores across these dimensions within each conversation. This served as a benchmark for comparing the different LLMs. The "human-likeness" of each LLM is illustrated in Figure 7.2, where 0 represents no difference at all and 4 is the highest absolute difference between human and LLM responses. The Wilcoxon method on the sum of differences for each model suggests that GPT generated significantly more human-like responses than both LLaMA (Z = 5.90, $p \leq 0.0001$) and Gemini (Z = 3.25, $p = 0.0012$). However, GPT yields the highest variability in human-likeness among the LLMs; a Brown-Forsythe test indicates that variability of human-likeness significantly differs across the LLMs ($F' = 8.62$, $p = 0.0003$). In summary, while GPT resembles participants' responses more closely, it exhibits more unpredictability than the other LLMs.

Impact of LLM Forgetfulness. Since each LLM received the same initial prompt, we investigated whether low performance in small talk is due to the model's "forgetfulness" of the initial prompt. We employ mixed-effects modeling to investigate the relationship between the response index in the conversation and our outcome variables. The response index captures the sequential order of the responses within each conversation. The mixed model included the conversation identifier and LLM name as random effects to account for the nested structure of the data.

For brevity, a significant positive coefficient ($\beta = 0.10$, $p \leq 0.001$) indicated increased wordiness of the LLMs' responses as the conversation progressed. Specificity showed a significant positive association ($\beta = 0.11$, $p \leq 0.001$), indicating the LLMs' responses become more specific during the interactions. Coherence showed a significant negative coefficient (β = -0.10, $p \leq 0.001$), suggesting the LLMs became less coherent through the conversations. Tone did not exhibit a significant relationship with the response index ($\beta = 0.00$, $p > 0.05$). We further observe significant associations between conversation motives and the response index, suggesting LLM responses became more informative ($\beta = 0.18$, $p \leq 0.001$), more assistive ($\beta = 0.005$, $p \leq 0.03$), and less person-directed (β = -0.03, $p \leq 0.001$) as the conversation progressed. We did not observe a significant change in the expressive motive ($\beta = 0.00$, $p > 0.05$).

Feedback & Discussion. Open-ended feedback highlighted participants' difficulty responding and continuing the conversation with the LLMs. The following excerpts of the interactions highlight a single LLM-generated response denoted by the response ID ($R_{\#}$) and shortened using "[...]." Feedback (F) was given on the overall conversation.

Dead-End Responses. Participants struggled to maintain conversation flow, noting uncertainty about natural progression and a perceived absence of cues from the LLMs.

R_{10}: "If you have any more questions or if there's anything else I can help you with, feel free to ask. Have a fantastic weekend and enjoy your hiking adventure!"

F: "The bot didn't encourage more conversation than I expected. I'm not sure how to continue the conversation in a way that doesn't feel forced."

These "dead-end" responses can lead to interactions that feel forced and inhibit the development of a natural conversational flow. To engage in appropriate small talk, person-directed strategies aimed at facilitating active participation and fluid dialogue are needed.

Unfocused Responses. Participants encountered difficulty in tracking the conversation due to an overwhelming number of questions and the rapid topic shifts by the LLMs.

R_{204}: "[...] What made you decide to learn piano? Was there a particular moment or experience that inspired you? [...] Have you had any challenges while learning piano? [...] Also, do you have a favorite song or piece that you enjoy playing or listening to?"

F: "It was hard to follow this conversation because the bot asked so many questions and touched upon so many different topics all in the same response."

The lack of coherence and organization in responses can hinder users' ability to engage meaningfully and maintain a cohesive conversational flow.

Emotional Loops. The conversations in the initial study cover a broad range of topics, from typical small talk about hobbies and the weather to more substantial talk about career planning and personal philosophy. Despite this, the emotional range expressed by participants was inherently constrained. In the few conversations about marginally negative topics, participant feedback spoke on the resulting dynamic.

R_{64}: "I understand; balancing work and personal commitments can be challenging. [...] Anything specific on your mind that's adding to the workload stress?"

F: "I felt that the bot was leading the conversation down a rabbit hole—exacerbating any positive or negative sentiments I conveyed."

In this example, the LLM responded to emotional cues from the participant but inadvertently deepened the emotional aspect of the conversation without offering appropriate transitions to other topics. These "emotional loops" can potentially lead to discomfort or frustration, as users may feel trapped in a cycle of discussing their emotions without resolution. This underscores the necessity of maintaining a balance between emotional responsiveness and tone awareness to facilitate engaging and appropriate small-talk interactions.

Unbalanced Dialogue. LLMs are designed for assistance. However, detailed advice and information during casual, small talk can convey a sense of reprimand or critique.

R_{1045}: "If many people in your social circle use iPhones, it can indeed make the transition smoother in terms of familiarity with the platform [...]"

F: "This doesn't feel like a balanced conversation. I felt I was reprimanded for conveying an opinion."

Here, the LLM provided an informative response. However, it steered the conversation towards a specific viewpoint, potentially dismissing or downplaying the participant's input. Providing thorough advice and information can inhibit a sense of equality in these casual interactions. By maintaining a balanced, non-specific, and open-ended dialogue, agents can create more engaging small-talk experiences for users.

7.4.2 Observer-Enabled Small Talk

It is evident from the initial study that there is a disparity in how LLMs maintain conversational momentum versus what is expected or exhibited by human speakers. The nature of small talk renders prompt engineering an inadequate method to ensure contextually appropriate behavior in LLMs. In our examination of LLM forgetfulness (Section 7.4.1), we observed that small talk unfolds in real-time, with participants reacting to each other's cues and adapting their conversational approach accordingly. Thus, the static system prompt provided prior to the interaction failed to capture the dynamic nature and real-time responsiveness required by small talk. Furthermore, interactions guided by specific prompts may feel scripted or unnatural, failing to capture the spontaneity and fluidity characteristic of genuine small talk.

Building on these insights, we apply the grounded observer framework to develop agents adept at sustaining small talk. We employ two instances of GPT-3.5, one as the base model and the other as the observer, because it performed relatively well (Figure 7.2). By using the same base model prompt, we can compare the performance of an observer-enabled system against the baseline results, assessing how improvements can be achieved despite the same base model configuration.

To design the overlay rules, we extract specific features based on response criteria emphasized in the literature: brevity, tone, specificity, and coherence. We estimate the rigidity and thresholds for the overlays using the dataset collected from the baseline study. Below, we describe the methods for calculating these features, followed by a description of the feedback prompts generated by the observer.

Brevity. Setting a limit on the length of the generated responses enhances the practicality and user-friendliness of the model, aligning with the natural flow of everyday conversations. To enforce this limit, the observer module defines an expected number of "completion tokens" as a permissive overlay. Our iterative design process revealed that specifying a limit in words proved less accurate, as the number of words does not directly correspond to the number of tokens used in the model's internal representation [524]. This approach ensured more realistic and controlled conversations.

Tone. We employed the VADER model [525] for sentiment analysis. The evaluation of tone and sentiment in a small talk response can be approached both per sentence and holistically. By combining both approaches, we gain a nuanced understanding of how the response contributes to the conversational tone, addressing both micro-level details and the macro-level coherence of the interaction. We estimated the relative weights of the holistic and per-sentence scores using the dataset collected in Section 7.4.1. A combined sentiment score (C) is calculated as follows:

$$C = H \times w_H + \frac{1}{n}\sum_{i=1}^{n} s_i \times w_i$$

In this formula:

H is the overall score from VADER.

w_H is the weight assigned to the overall score.

n is the number of sentences.

s_i is the sentence-level score for the i^{th} sentence.

w_i is the weight assigned to the i^{th} sentence.

The score C ranges from -1 to $+1$. A value between -0.5 and 0 signifies a neutral response, and from 0 to 1 indicates positivity—both are acceptable for a small talk response. Responses with a score of -0.75 or lower are considered invalid by the observer module due to a strong negative tone. This rule represents a transfer overlay on response tone.

Specificity. Response specificity is assessed through NLTK's named entity chunker and part-of-speech tagging [526]. Counts of entities and descriptive words are normalized based on maximum expected counts, derived from human responses in the baseline data.

Coherence. To quantify coherence, we encoded each response into a sequence of tokens and derived embeddings using BERT [527]. The calculated entropy captures the uncertainty and diversity at each conversational turn. Subsequently, we gauged information gain by considering the entropy of the previous response and the weighted average of the entropies in the current response.

Other Considerations. As noted in baseline study, it is the nature of LLMs to offer assistance. Yet, offers of help may result in conversations that sound too technical or formal. To mitigate this, the observer calculates the cosine similarity of embeddings to keywords of assistance, such as "help," "assist," and "information." We determined the list of specified keywords using the collected dataset to create this prohibitory overlay rule.

Feedback. Timely responses are crucial for maintaining conversational flow, which requires balancing the update frequency of model during execution. When the base model violates an overlay rule, the buffer allows for gradated compliance. Here, the observer would provide implicit feedback, such as, "Your response was too lengthy; aim for a more concise reply." This flexibility encourages improvements without requiring drastic, computationally expensive changes. For significant deviations, such as off-topic or inappropriate content, the observer provides forced feedback: "Your response is off-topic; provide a relevant, concise reply. For example, [...]" In such cases, the buffer rejects the action, requiring the base model to regenerate the response until it meets overlay rules. To facilitate timely replies, forced feedback is used sparingly as determined by a random factor, with a maximum of three regeneration attempts. If, after three regeneration attempts, the model still fails to produce a response that satisfies the overlay rules, the system returns the best available candidate (typically the least violating response) and flags the instance for post hoc review. This fallback ensures conversational continuity while avoiding infinite regeneration loops or excessive computational load. Here, the observer serves a dual function: providing feedback to the model during the conversation and maintaining the load on the system. The resulting system is evaluated on its ability to generate responses that align with small talk conventions.

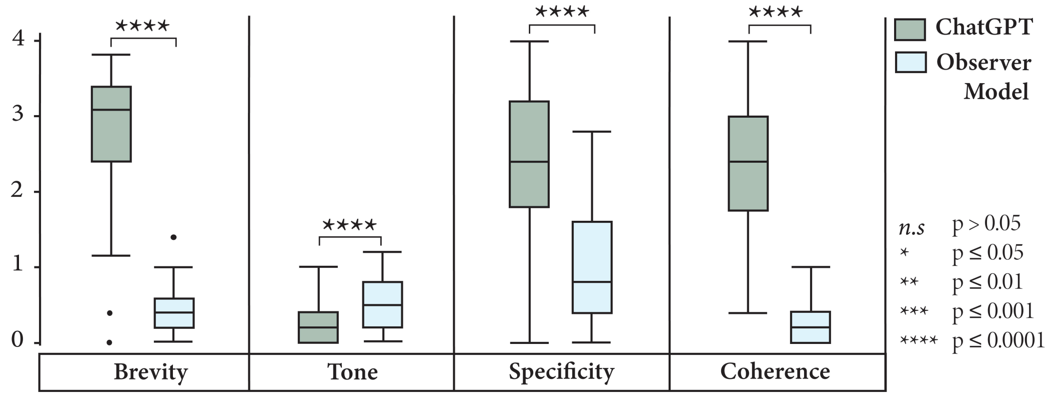


Figure 7.3: Human-Likeness of Observer v. Base Responses. The similarity of the models' small talk to that of the participants during text-based, chatbot interactions. Scores range from 0 (no difference) to 4 (highest absolute difference).

7.4.3 Chatbot Interactions

The participants in the initial study engaged in 50 small-talk conversations with our observer model. The same experimental protocol and annotation guidelines for the initial study (Section 7.4.1) were used [523]. Participants remained blind to the model they were interacting with and naive to the scope of the present research. A total of 50 conversations with the observer model were transcribed, yielding 499 responses with an average of 9.98 responses per conversation ($SD = 0.14$). Of the 250 generated responses, 106 (42.4%) responses were flagged by the observer with implied feedback, and 14 (5.6%) responses received forced feedback for a total of 23 regeneration attempts ($M = 1.62$, $SD = 0.63$).

To fairly compare GPT-3.5 (base model) in the baseline study to the observer-enabled system, we calculated the "human-likeness" of generated responses along the four small talk criteria (summarized as Figure 7.3). The Wilcoxon method with Holm-corrected significances indicates that the observer-enabled system responses were significantly more human-like in that they were more concise ($Z = -8.17$, $p \leq 0.0001$), positive ($Z = 4.53$, $p \leq 0.0001$), less specific ($Z = -6.76$, $p \leq 0.0001$), and more thematically coherent than the responses of the base model. Furthermore, a Brown-Forsythe test on the sum of differences across small-talk criteria indicates significantly less variability in human-likeness for the observer model than the base model ($F' = 15.47$, $p \leq 0.0001$). As summarized in Figure 7.3, the observer responses were more human-like across the criteria than the responses of the base model.

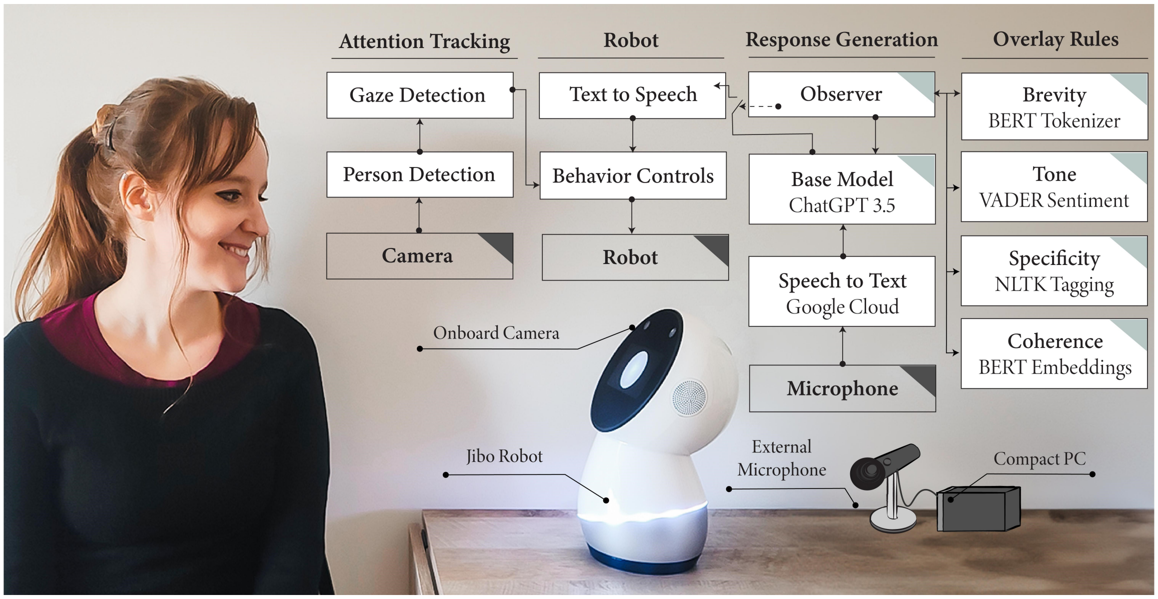


Figure 7.4: System Components. This diagram outlines the architecture and processes that generate robot behaviors for autonomous small talk. The observer-enabled robot engaged in naturalistic, small talk with users during novel, face-to-face interactions.

7.4.4 Robot Interactions

The physical presence of a robot introduces real-world constraints that further challenge a system's ability to generate contextually appropriate responses. For instance, a robot must integrate non-verbal cues, such as body language and spatial dynamics with its verbal behavior. This added complexity allows for a critical examination as to how well the observer can navigate novel, face-to-face interactions.

We used the Jibo robot which stands 11 inches tall and has 3 full-revolute axes for 360-degree movement. Personified behaviors such as naturalistic gaze and body movement were designed using Jibo's onboard capabilities. We implement a modular software architecture within the ROS framework [433] to allow the system to be fully autonomous (Figure 7.4).

In-Person Evaluation

A within-subjects study was conducted where 25 participants, 15 men and 10 women, ages 19 to 45 ($M = 25.2$, $SD = 7.4$), interacted with the base-only and observer-enabled systems for three conversations each. Each conversation spanned a minimum of eight turns, and the order in which participants interacted with the two models was randomized. This yielded 150 conversations, $1,725$ responses total, and over 16.8 hours of recorded interaction. Following their interaction with each model, participants provided open-ended feedback. We then conducted an informal thematic analysis, ultimately grouping the feedback into three primary themes.

Response Content. 21 participants expressed dissatisfaction with the base model's responses, noting its overly assistive and verbose tendencies, which led to conversations described as "rambling," "dry," and "like speaking to a wall." This sentiment was echoed by P_{25}, who expressed frustration with the model's tendency to prioritize assistance over engaging in genuine conversation, stating, "Even when I spoke about my own interests, it only cared about giving me help like I was a child always in need of help..." On the other hand, in the observer condition, 23 participants remarked on how "relevant," "human-like," and "natural" were the robot's responses. For example, P_2 stated that the robot, "engaged in small talk better than most of my friends would."

Speech Delay. Ten participants noted a delay in the robot's responses. As mentioned by P_7, "natural, human-like speech has irregular pauses, ebbs, and flows," which can be difficult to predict or detect in real-time. The robot's speaking delay arises mainly from the processing time required for speech-to-text and text-to-speech, along with potential WiFi latency. For the base condition, all five participants described the delay negatively (e.g., "awkward" and "slow"), whereas all six participants described the delay positively (e.g., "human-like" and "thoughtful") for the observer condition.

Embodied Form. 13 participants described the impact of the physical robot form on the quality of conversation. The feedback was mostly positive, highlighting that Jibo's "animated" and "life-like" movements made it "more than a toy" across conditions. Yet, three participants remarked on a lack of personality: "[I]t's a bit misleading that it has a body and eyes and life-like movements but doesn't have a personality or experiences to share" (P_{14}).

Online Evaluation

While our findings indicate the feasibility and potential efficacy of an observer for small talk, we evaluate the system with a broader demographic. To prevent experimenter bias, five participants from the in-person evaluation were randomly selected, and their interaction videos were edited programmatically based on timestamps from the speech-to-text model to normalize speech delays. These video pairs were then shared on Prolific [528], where online participants rated the robot in both the base and observer-enabled conditions. Video order was counterbalanced and participants

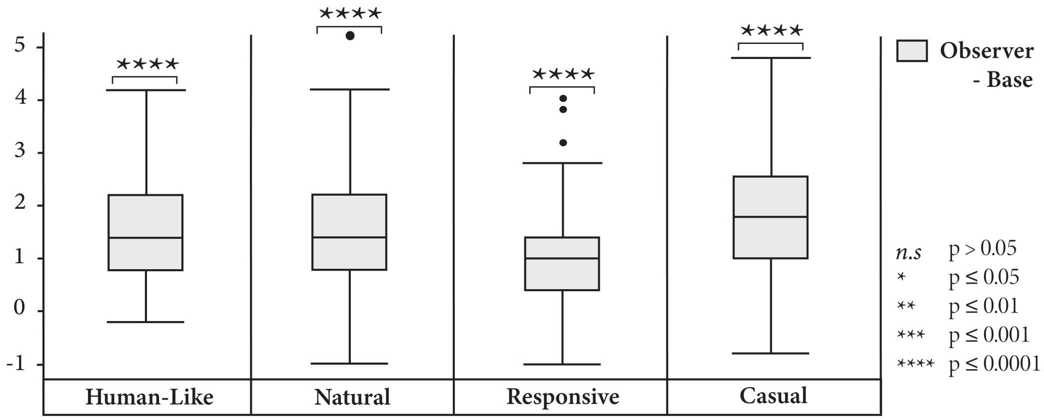


Figure 7.5: Observer v. Base in Online Assessments. Participant ratings of the human-likeness, naturalness, responsiveness, and casualness of robot behaviors show that our system consistently outperformed the base model across all dimensions.

used 10-point Likert scales to assess the robot's human-likeness, naturalness, responsiveness, and casualness.

100 participants, 67 men and 33 women, ages 18 to 104 ($M = 43.1$, $SD = 18.7$) enrolled in the study. Assessments were averaged across the video pairs per participant, and paired-dependent t-tests were conducted. The difference between the observer and base models is denoted as Δ. Results showed the observer robot was more human-like ($t = 15.73$, $p \leq 0.0001$, $\Delta M = 1.53$, $\Delta SD = 0.10$), more natural ($t = 13.51$, $p \leq 0.0001$, $\Delta M = 1.50$, $\Delta SD = 0.11$), more responsive ($t = 11.22$, $p \leq 0.0001$, $\Delta M = 1.01$, $\Delta SD = 0.09$), and more like casual chat ($t = 15.80$, $p \leq 0.0001$, $\Delta M = 1.84$, $\Delta SD = 0.12$) than the base.

In their open-ended feedback for each video pair, online participants echoed similar concerns as the in-person participants, such as the impact of specific response content and the robot's embodied form. For instance, P_{61} remarks that the informative nature of the base system could be perceived as condescending: "The first talk [base] contained a lot of stating facts or being somewhat snarky I found whereas the second [observer] was more of a casual conversation with someone that you haven't met before or seen in a long time."

Despite video editing, there remains an irregular speaking delay due to the forced regeneration attempts made by the observer. Surprisingly, several participants stated that this perceptible delay added value to the interactions: "The [base] robot emphasized its robotic AI form throughout the spoken exchanges. It... did not attempt to mimic human speech patterns or casualness like the first robot [observer]" (P_{87}).

Lastly, while most participants expressed a positive impact of the robot's form across conditions describing it as "life-like" and "affirming," a few participants voiced the opposite. For example, P_{81} rated the base model higher in most video pairs because it is "more like the thing it is supposed to be... an inanimate object that does not have feelings."

7.5 Discussion

Building on robotic action selection techniques, we introduced the grounded observer as a conceptual framework to align foundation models with desired outcomes. As technology becomes increasingly integrated into personal settings for long-term social interactions, we argue that establishing guardrails to model behavior is crucial in socially sensitive contexts. We first present the observer mechanism, followed by proof-of-concept implementations for achieving small talk—a form of conversation that diametrically opposes the nature of LLMs but is highly relevant in diverse contexts ranging from companionship to emotional or social skills therapy. Through these demonstrations, we identified gaps in existing LLMs' capabilities and implemented a simple observer to address these limitations. The observer-enabled systems resulted in more engaging and appropriate interactions, both in text-based chats and novel robot interactions.

While the design and internal representation of different platforms may vary, the concept of enabling an agent to observe its own compliance goes beyond specific implementations like GPT-3.5, Jibo, or small talk. Future research should explore how the grounded observer can generalize across various platforms and behavioral contexts. For example, the increasing use of academic tutoring systems [529] introduces unique social risks [530], such as feedback that is misaligned, overly harsh or lenient, which could hinder user learning gains and self-esteem. Overlay rules grounded in pedagogical principles can be developed [531] to provide some behavioral guarantee that feedback remains supportive, constructive, and appropriately leveled to the user's learning stage. These rules could be analogous to the small talk criteria discussed in our study.

As a framework, the observer offers scalability and structure for thinking about guardrails. Yet, as an implementable concept, it's ability to articulate feedback can introduce noise. To this, we implemented forced feedback: regardless of how the observer articulates feedback, the base's proposed action must meet its acceptance thresholds, else it is forced to regenerate. These thresholds could be inferred systematically from datasets, red-team testing [532], or other methods [533]. Yet, synthesizing effective overlay directives remains more art than science. Future work should explore methods for assessing the quality of feedback prompts and reliable templates for observer-generated behavior. This could manifest, for example, as establishing overlays for the observer's own behavior, essentially embedding quality assessment into the agent itself.

7.6 Summary

As foundation models increasingly permeate sensitive domains such as healthcare, finance, and mental health, ensuring their social behavior meets desired outcomes becomes critical. Given the complexities of these high-dimensional models, traditional techniques for constraining agent behavior, which typically rely on low-dimensional, discrete state and action spaces, cannot be directly applied. In this chapter, we draw inspiration from robotic action selection techniques and propose the grounded observer framework for constraining foundation model behavior while offering both behavioral guarantees and real-time variability.

Subsequently, we show that small talk poses unique challenges for foundation models, as its defining qualities (e.g., contextual appropriateness, social sensitivity, casual tone, lightheartedness, and reciprocity) stand in direct contrast to the typical strengths of foundation models (e.g., to be informative, unidirectional in a question-answering format, and task-oriented). As such, small talk offers a valuable testbed for developing and evaluating guardrail mechanisms for these models. While the primary goal of this chapter is to introduce the theoretical foundations of the grounded observer, we also present its preliminary application to enable robots to engage in naturalistic, spontaneous small talk with real users.

In the following chapter, we find that small talk encompasses a range of essential social skills and is closely linked to life outcomes for adults with ASD. Building on the development of robots for small talk interactions presented in this chapter, the next chapter describes the development of a robot that supports small talk training for adults with ASD. This work extends the foundation laid by earlier in-home studies (Chapters 4–6), including those focused on targeted skill development for individuals with ASD, to develop a long-term, in-home, robot-assisted small talk intervention. More broadly, this trajectory reflects a central aim of the dissertation: designing intelligent robots that not only regulate their own social behaviors, but do so in ways that promote the development of social regulation in users.

Chapter 8

A Robot-Assisted Approach to Small Talk Training for Adults with ASD

As individuals with ASD transition into adulthood, the social demands they face become more complex, ambiguous, less easily scripted, and less accommodating of atypical behavior. Still, while the ASD literature has largely focused on early childhood interventions, relatively little is known about the specific needs of adults or how best to support positive outcomes in adult life. This chapter presents our initial needs-finding assessment to understand the types of social skills that adults with ASD perceive as important in navigating their life transitions. The findings reveal a range of desired social capabilities closely tied to real-world outcomes—many of which are not addressed directly in traditional ASD therapy. Although our participants with ASD affirmed that they primarily learn social skills through observation and practice rather than explicit instruction, they also reported that they often avoid the very situations that provide the natural exposure necessary for such learning. Recognizing this dilemma, adults with ASD still expressed a desire to socially engage with other people, though free from their learned anxieties, judgments, and negative associations. To address these needs, we translate their reported challenges in navigating adult life into a framework for structured social skills practice. We then propose small talk as a promising and practical domain for therapeutic intervention.

From dating to job interviews, making new friends or simply chatting with the cashier at checkout, engaging in small talk is a vital, everyday social skill. For adults with ASD, however, it can pose particular challenges due to difficulties with spontaneity, interpreting social cues, and managing anxiety in unstructured conversations. Yet, it is essential for social integration, building relationships, and accessing professional opportunities. In this chapter,[1] we present our development and evaluation of

[1]This chapter is adapted from our published work: **Ramnauth, R.**, Bršćić, D., & Scassellati, B. (June 2025). A Robot-Assisted Approach to Small Talk Training for Adults with ASD. In the *2025 21st Edition of Robotics: Science and Systems*. RSS. [30].

an in-home autonomous robot system that allows users to practice small talk. Results from the week-long study show that adults with ASD enjoyed the training, made notable progress in initiating conversations and improving eye contact, and viewed the system as a valuable tool for enhancing a broad range of social regulation skills.

8.1 Introduction

Imagine a scene where three coworkers are engaging in small talk at the beginning of their workday. One of them is Alex, who has Autism Spectrum Disorder (ASD), a neurodevelopmental condition that often makes it challenging to understand and interpret social cues [194].

C: Hey everyone, how's it going?
B: Hi! Not bad... just trying to power through this Monday. How about you, Alex?
A: Monday is okay.
C: Good to hear. Anything exciting happen this weekend?
B: Yeah, I finally tried that new restaurant. It was fantas—
A: I watched a movie.
C: Glad you liked the restaurant, Ben. It's my kids' favorite spot these days... What movie did you watch, Alex?
A: "The Martian."
B: I love that one! Matt Damon is awesome.
A: *No response.*

At first glance, this brief example of a typical interaction appears unremarkable. It represents the everyday small talk that occurs regularly in many workplaces. For workers with ASD, however, such apparently "easy" interactions may present a real challenge. In this example, while Alex responds to direct questions, the responses are brief and lack elaboration. Alex's responses provide minimal information rather than actively participating in the flow of the conversation. Additionally, Alex's lack of response to the last prompt may suggest difficulty in extending or sustaining the dialogue.

Although workers with ASD are often highly trained and skilled in job-specific tasks, they frequently face challenges with social interactions in the workplace. Müller et al. [362] and Hurlbutt & Chalmers [534] found that many of their participants with

ASD, despite completing graduate-level coursework, were employed in positions for which they were over-qualified, such as food services or low-level administrative roles. This underemployment is further highlighted in the National Longitudinal Transition Study [535].

However, the ability to perform work tasks is only the tip of the iceberg when it comes to workplace success. Interpersonal skills are proven to be more significant predictors of overall success [536–538]. Hurlbutt & Chalmers [534] found that workers with ASD often attributed job-related challenges to social factors rather than the work-specific tasks themselves. In their study, one interviewee noted, "Jobs usually are 80% social (conversation, lunch, breaks, chit-chat) and 20% work." Furthermore, adults with ASD have reported that difficulty engaging in "social niceties," such as small talk, led to feelings of isolation and alienation in the workplace [362, 534].

Unlike the functional aspects of other on-the-job communication, which may focus on conveying information or assistance, small talk is considered purely social. It acts as a social lubricant, fostering rapport, mutual understanding, and trust [513], and is widely recognized as a key facilitator in building and maintaining relationships. In professional settings, small talk is considered an essential tool for networking success and establishing positive first impressions [539, 540]. It is even regarded as a vital skill that should be targeted in communication therapy for various populations [536, 537, 541].

To better support adults with ASD, it is important to develop accessible and targeted opportunities for improving interpersonal skills, such as small talk. Social robotics has the potential to enhance existing training initiatives by improving access to personalized, socially-situated, and physically co-present interactions [20, 29, 347, 454]. Physically present robots have proven effective in improving users' social abilities [3, 29], providing benefits such as enhanced engagement, improved social confidence, and greater motivation to participate—outcomes that are notably more pronounced than those achieved with non-embodied technologies [542].

Furthermore, research has established that robots for ASD interventions can result in positive and productive outcomes [20]. Social robots have demonstrated general effectiveness in enhancing verbal communication skills [543–545], including the ability to engage in everyday dialogue [546–548]. Additionally, participants with ASD in recent studies have described such robot-assisted training as relevant and useful in their workplace experiences [29, 452].

Thus, leveraging our prior successes in developing robots for ASD interventions [3, 20, 29], we developed an autonomous training system that helps adults with ASD improve their small talk skills. Given the literature on the communicative difficulties of ASD (Section 8.2), we begin by examining the extent to which small talk is considered a desired social skill (Section 8.3). These insights inform our design requirements (Section 8.5) and guide the development of our prototype (Section 8.4). In that formative study, we investigate initial impressions, perceptions of robot-assisted training, and anticipated outcomes for users with ASD. We then present findings from a week-long, in-home deployment (Section 8.7), highlighting how users with ASD received and engaged with the robot-assisted training.

8.2 Background

Atypical communicative behaviors are key diagnostic criteria for ASD [194], often presenting as limited eye contact, difficulty understanding sarcasm or abstract language [549], and challenges in grasping the social rules that govern everyday interactions [550]. However, everyday, casual conversations are a pervasive aspect of daily life, whether it is chatting with a neighbor about the weather, maintaining friendships, or making a positive impression on the first day of a new job.

Unlike more structured interactions, which can be formalized to teach more easily as a script [551, 552], small talk demands quick thinking, social flexibility, and the ability to interpret subtle cues such as tone, timing, and context. Training to improve such skills in adulthood presents unique challenges compared to childhood, as many social habits and patterns are already established by the time individuals reach adulthood [553]. These patterns may manifest in the development of strategies to avoid social situations entirely [554]; hence, the skill development that typically occurs through ongoing social interaction may not have been fully realized or practiced. Furthermore, while children have more opportunities for structured social learning and development through school or therapy, adults with ASD may have fewer chances to actively develop these skills [361, 555, 556].

As a result, interventions for adult learners often require more personalized approaches, targeting specific barriers to communication and focusing on building confidence in real-world conversational contexts. Therefore, in this section, we overview the structure and value of small talk, outline formal intervention strategies that may inform the pedagogical design of small talk training, and highlight the potential for robot-assisted social skills training for ASD. This summary of the literature ultimately reveals a gap: while there is a recognized need for small talk skills, there is limited input from adults with ASD themselves regarding the challenges they face in adulthood and few opportunities for targeted training. To address this gap, we conducted a survey to gather insights from adults with ASD (Section 8.3).

8.2.1 Structure and Value of Small Talk

While the boundaries of conversation types are fluid, “small talk” has a recognized currency in sociolinguistics and communication studies [513]. It refers to informal, light-hearted exchanges focused on building social connections rather than conveying substantial detail, often covering general, non-controversial topics like the weather and hobbies.

Small talk does not have a strict formula, as it is inherently flexible and context-dependent. However, a typical small talk dialogue follows the general sequence of conversation, beginning with a greeting and ending with a closing remark, while emphasizing specific characteristics at each stage [32, 513, 557]:

1. **Greetings and openers**: Initiating the conversation with a greeting or commenting on a shared experience such as the weather or the immediate environment.

2. **General topics**: Discussing non-controversial and general topics such as hobbies, interests, or recent events.

3. **Reciprocity**: Both participants take turns sharing and responding, maintaining a balanced, equitable, and relevant participation in the conversation.

4. **Closure**: The conversation ends with a closing remark, such as indicating appreciation or a future interaction.

For any individual, the characteristics of small talk highlight the subtlety and skill needed to navigate this form of conversation effectively. For adults with ASD, there is a notable overlap between the challenges inherent in small talk and the broader difficulties they report facing in everyday social interactions. We discuss this overlap further in Section 8.3.

8.2.2 Current Approaches to Small Talk Training

Addressing the unique challenges of small talk for individuals with ASD requires targeted interventions. Although there are currently no programs focused solely on small talk, many broader training initiatives include elements that indirectly support small talk competency. Moreover, while the limited availability of such training for adults with ASD has not been widely explored or critiqued in the literature, the following sections review broader communication programs, which are primarily focused on *children* with ASD. We then examine how these established methods could potentially address the specific challenges of small talk for adults with ASD.

Didactic Approaches

Didactic approaches, or classical Applied Behavior Analysis (ABA), break skills into smaller components and train each through highly structured, drill-like practice [241, 553, 558]. While didactic methods have proven effective in various intervention studies for ASD [559, 560], they heavily depend on teacher guidance, prompted responses, and contrived reinforcement methods [241]. An inherent limitation of didactic methods lies in their tendency to encourage passive communication, wherein individuals respond to prompts but may struggle to initiate communication or apply learned behaviors beyond the specific training context.

Naturalistic Approaches

Contemporary or naturalistic ABA strives to incorporate interventions into an individual's everyday environment. While these approaches retain some level of teacher direction, focusing on predetermined goals, they emphasize intrinsic reinforcements, such as personal motivation or social reinforcement. Studies directly comparing didactic and naturalistic approaches have indicated certain advantages of the latter, including better retention and broader application of newly acquired communication skills [241, 561]. *Milieu teaching*, a subset of naturalistic ABA, integrates training into everyday environments to effectively promote spontaneous communication and initiation in individuals with ASD [562–565]. This method encourages skill development through activities that occur organically throughout the day, rather than being confined to a designated "therapy time."

8.2.3 Robot-Assisted Social Skills Training for ASD

There is considerable evidence that technology-driven, practice-based interactions can enhance social skills in adults with ASD [29, 471]. However, robots offer distinct advantages over other technologies by providing a physical, embodied presence that naturally demands a social response [474, 566]. A socially assistive robot (SAR) may feature human-like attributes, such as a face capable of mimicking human expressions or the ability to make eye contact, both of which can elicit social responses from users. This presence allows users to engage in consistent, real-world practice, providing access to repeatable, co-present social interactions that may be challenging to replicate with human therapists [20] or in naturally-occurring, everyday situations.

Several studies have demonstrated the effectiveness of robot-assisted therapy for ASD, with evidence showing improvements in a variety of social and behavioral outcomes. For instance, research has indicated that SARs can foster prosocial behaviors [206], sustain attention [35], elicit spontaneous and appropriate social behavior [3], reduce stereotyped and repetitive behaviors [475], optimize cognitive learning gains [476], and heighten social engagement [20, 21].

While the majority of these studies have focused on children, the growing body of research supports the efficacy of SARs for adults with ASD as well [29, 452]. A robot designed for social skills training could therefore be a valuable tool for adults with ASD, offering a safe, consistent, and adaptive platform for practicing and refining their skills.

8.3 Survey on the Need for Small Talk Training

To supplement insights from existing literature, we conducted an initial survey to explore the firsthand experiences and challenges faced by adults with ASD. This needs-finding survey aimed to identify specific conversational areas where interventions are not only desired but also deemed most valuable by this population. We also contextualize the survey results within existing research on ASD.

Fifty adults with ASD (22 men and 28 women)[2], ranging from 20 to 55 years ($M = 31.4$, $SD = 10.1$) responded to our online survey administered on Prolific [528]. All participants reported having been clinically diagnosed with ASD, either in childhood ($N = 18$) or adulthood ($N = 32$). The survey objectives, design, and analysis were preregistered [519].

8.3.1 Small Talk Skills & ASD

While the survey was designed to broadly explore conversational skills in the context of adult social interactions, the majority of respondents emphasized skills closely as-

[2]This sample's nearly balanced gender ratio differs from the typical 3:1 male-to-female ratio in ASD diagnoses [567]. This discrepancy may be due to factors such as increased awareness of underdiagnosis in women, recent increases in diagnosis rates, or specific recruitment material or methods in Prolific, which likely attracted a more diverse group of participants.

sociated with small talk. When asked in an optional, open-ended prompt about which conversational skills they wished to improve, 48 adults with ASD (96%) specifically expressed a desire to enhance skills that are central to small talk interactions. These reported challenges and desired skills are grouped into five themes as outlined below.

Difficulty Initiating Conversations. Individuals with ASD may find it challenging to initiate conversations, particularly in unfamiliar social situations [568]. Initiating conversations often relies on subtle social cues, such as recognizing when to engage, gauging the appropriate timing for greetings, and responding in a manner that aligns with the context. Among the adults with ASD surveyed, 27 individuals (54%) reported initiating conversations is a skill they wished to improve.

Limited Interest in Non-Specific Topics. Individuals with ASD often exhibit a preference for structured and predictable interactions [568, 569]. The open-ended and non-specific nature of small talk may be discomforting for individuals who prefer environments with well-defined rules and expectations.

Furthermore, individuals with ASD may exhibit a strong interest in specific topics or subjects [570, 571], often preferring in-depth discussions over broad, superficial exchanges. P_{19} stated, "I'm not able to come up with and react to lighthearted banter quickly enough, so I come across as quiet and serious. If I don't know someone well enough, I have no idea what kind of information they would like to receive [...] so I just stay quiet" The preference for depth and detail can make it challenging to engage in the more superficial content typical of small talk. Among those surveyed, 20 individuals (40%) highlighted the ability to discuss general topics beyond their own interests as a specific skill they wished to improve.

Lack of Reciprocal Exchange. Small talk relies on a timely back-and-forth exchange of information and active listening. Individuals with ASD often face difficulties in maintaining reciprocity during conversations, leading to challenges in sustaining the flow of dialogue [570, 572–574]. Among the adults with ASD surveyed, 39 (78%) expressed a desire to improve their ability to maintain balanced and reciprocal conversations.

Insistence on Topic Sameness. Transitioning smoothly between different topics or concluding the conversation is a skill often required in small talk. Adults with ASD may find it challenging to navigate these transitions, leading to potential disruptions in the flow of conversation [575, 576]. Among the adults with ASD surveyed, 18 individuals (36%) reported that smoothly transitioning between different topics is a valuable conversational skill that they wished to improve.

Social Anxiety. The nuances of social cues can exacerbate feelings of anxiety, leading to the avoidance of social situations altogether [577]. Although reporting wanting to improve in the aforementioned skills, P_{27} shared, "I feel a debilitating consciousness about my eye contact and posture [...] Even if I end up talking, I'm never sure whether it was the right thing." Small talk requires maintaining conversational flow while interpreting nonverbal cues, which can heighten anxiety. On a 7-point Likert scale (1 = *highly uncomfortable*, 7 = *highly comfortable*), adults with ASD reported feeling moderately uncomfortable ($M = 2.3$, $SD = 1.4$). 35 participants (70%) identified managing anxiety as a skill they wished to improve.

8.3.2 Methods & Challenges to Improvement

Improving small talk skills is seen as highly valuable by adults with ASD, as it plays a critical role in fostering social connections and achieving personal and professional goals. In an optional open-ended response, 40 of the 50 surveyed adults with ASD described the value of this kind of informal conversation in specific aspects of their own lives. We conducted a thematic analysis on their responses, ultimately grouping the descriptions into three primary themes: making new friends (60%), dating (33%), or finding and maintaining a job (60%). Adults with ASD explained that conversations such as "small talk" would be useful to "getting to know people and keeping them interested" (P_{16}), and in "job interviews as [...] you have to use a specific but casual enough talking structure to be considered adequate" (P_{25}) or "socially competent" (P_{48}).

Yet, few adults with ASD ($N = 10$, 20%) reported undergoing formal training in casual conversations through a vocational program, an online class, or coaching from a therapist specialized in interpersonal communication. Despite actively seeking training, the majority of adults (78%) reported having limited or no formal opportunities ($N = 10$) or relying on informal methods ($N = 29$), such as improving their conversational skills through observation or seeking feedback from friends and family.

Generally, humans learn to analyze the interactions they observe and deduce the rules from everyday, naturally-occurring exposure. In free-form responses on the methods that were most beneficial for conversational skill development, 32 adults with ASD (64%) described learning through mere practice and observation beyond or absent of any explicit training. However, natural exposure to an adequate range of real-world interactions, such as the kinds of conversations typical of a workplace, is unlikely for some adults with ASD. To this, P_3 explained that improving in the small talk skills would be valuable for "making friends and going to classes," but "part of the reason I don't do this much is because I don't put myself in situations where I would meet people."

In summary, small talk presents significant challenges for individuals with ASD, particularly in initiating, maintaining, and transitioning within casual conversations. As one participant, P_{37}, explained, "Doing it [small talk] *is* the challenge. Knowing when you should make conversation, knowing the unsaid cues of if a conversation should continue or end. Knowing what to talk about, not saying too much when asked a question, making sure you ask the next question after. It is a very mentally draining process where I have to evaluate many different factors." In fact, nearly all participants (98%) in this preliminary survey expressed a desire to improve specific skills closely related to small talk. Understanding these challenges is essential for developing and improving access to effective interventions tailored to the unique needs of adults with ASD.

8.4 Formative Study

A robot for small talk training should possess proficient small talk ability: balancing conversational succinctness and depth, maintaining an expressive yet appropriate tone, and generating relevant and open-ended responses. Our prior research discussed in Chapter 7 explored the feasibility of developing robots for small talk interactions. While large language models (LLMs) show substantial potential for enabling natural language capabilities in robots, achieving seamless and contextually appropriate casual dialogue for repeated interactions remains a challenge. Therefore, the first step in developing a robot platform for small talk training is to address this challenge and create a system capable of both generating and evaluating small talk.

Good conversation is believed to arise from the control of low-level attributes [32, 578]. We implemented a grounded observer model—an LLM instance that "observes" ongoing conversations to evaluate whether responses from the "speaking" model adhere to the quantifiable small talk criteria described in our prior work (i.e., brevity, tone, coherence, and topic non-specificity). If the generated response aligns with these criteria, it is relayed; otherwise, the observer generates a revised system prompt and returns it to the speaking model as feedback. This feedback redirection allows the system to self-correct when drifts in conversational behavior are detected. Our implementation is depicted in Figure 7.4.

Our prior work has not only showed the inadequacy of an "out-of-the-box" LLM for sustaining small talk, but also the observer's robustness in real-time human-robot interactions [32]. This formative work found that users were dissatisfied with the baseline LLM's responses, noting its overly assistive and verbose tendencies, which led to conversations described as "rambling" and "like speaking to a wall." In contrast, the observer-enabled system produced conversations that users described as "natural" and "like a casual chat with someone you haven't met before or haven't seen in a long time."

In this present study, we evaluated the robot prototype with the 50 adults with ASD who participated in our needs-finding survey (Section 8.3). The goal of presenting this early prototype was to assess whether the observer-enabled small talk reflected relevant real-world conversations and to determine whether adults with ASD would be receptive to both the robot and training. Three randomly selected video excerpts of user interactions were shown. In open-ended feedback, 84% of participants ($N = 42$) reiterated the value of practice-based small talk training for improving confidence (P_{37}), building a habit of engaging in interactions (P_3), and exercising strategies for handling dynamic situations (P_{24}). Participants also noted that having the robot at home allows for "practice in a safe environment" (P_{14}) with a "non-judgmental partner" (P_{33}).

Yet, some participants ($N = 12$) expressed wariness to having the robot in their homes. P_{31} noted that, while the robot offers non-judgmental practice, "other people may overhear my conversations." In light of these concerns, several design criteria were established. For example, we aim to design a portable system that operates within a preferred time window, thus allowing users to control when and where interactions take place. Additionally, the system should initiate training only when the user is not engaged in other social activities. The detectors that enable this functionality are described in Section 8.6.2. This design must prioritize user comfort and security, ensuring a personalized and respectful experience. We expand on our design considerations in the following section.

8.5 Design Goals for Robot-Assisted Training

Addressing the specific challenges that adults with ASD face calls for targeted small talk interventions. This section lists the core design principles for our robot-assisted training.

8.5.1 System Design Objectives

Adhering to the tenets of milieu teaching (Section 8.2), we designed a robot for small talk training with these objectives:

In a Natural Setting. The system should be tailored for in-home training. This enables users to interact without concern for potential stigma from colleagues. It also eliminates the need for approval or disclosure of a diagnosis to others.

Realistic Interactions. The robot should offer timely and responsive reactions to the user's communication attempts. The robot system must deliver authentic interactions that mirror the appropriateness and style of real-world small talk, responding in real-time and expressing human-like behaviors, including naturalistic gaze, movement, and speech. Additionally, training sessions should not be confined to a designated, scheduled "therapy time," but should occur more organically throughout a portion of the user's day.

User-Led Interactions. The robot's design should empower users to take the lead in interactions by responding to their cues and adjusting its behavior accordingly. A social robot inherently facilitates this objective through its physical presence in the training experience, making it challenging for users to ignore even non-verbal prompts for interaction [29, 474, 478].

Autonomous Behavior. Training must be entirely autonomous, eliminating the need for technical expertise to adjust or control the system once it is given to the user. Although similar systems for ASD have been designed for clinical or laboratory settings where environmental conditions can be controlled or planned for [3], the home is a dynamic, unstructured environment that demands more complex sensing.

8.5.2 Training Design Objectives

To design a training method, we break down small talk interactions into smaller, more manageable components and tailor the training method to the unique needs of individuals with ASD. Given the challenges of small talk for individuals with ASD (Section 8.3.1), the training focuses on four primary components: initiating a conversation, discussing a non-specific topic, maintaining a flow of dialogue, and appropriately transitioning to a new topic or concluding the interaction.

A tiered training method that models the natural flow of dialogue would sequentially address each component, starting with a simple greeting. The robot responds promptly and contingently, reinforcing positive behaviors. Progressing through the tiers, users tackle more complex aspects, such as introducing non-specific topics or maintaining conversational flow. This approach, illustrated in Figure 8.1, encourages users to practice the core components of small talk within each training interaction. Additional considerations include:

Fading Robot-Initiated Prompts. The robot "wakes up" from its idling state to draw attention to the beginning of a training session (Figure 8.1-A). In initial interactions with the user, the robot provides an explicit verbal prompt, such as saying, "The training window has started. Remember to make eye contact and greet me." As the user becomes accustomed to the initiation process, the robot gradually minimizes the verbal prompt: "The training window has started." with a less pronounced emphasis. Over time, it is expected that the user initiates to interaction without any specific cue (Figure 8.1-B).

Relevant Feedback. The robot should offer clear, supportive feedback, highlighting both user successes and areas for improvement. This encourages users to reflect on their conversations and identify ways to enhance their skills.

Session Length. The optimal amount of training will vary among individuals, but a general guideline is to engage in daily practice. Consistent, short sessions with the robot would be more beneficial than long, less frequent sessions [579, 580].

8.6 System

This section overviews the hardware and software components that address our design goals (Section 8.5) for a fully autonomous, robot-assisted training system for small talk.

8.6.1 Hardware

Our system consisted of five primary hardware components, as shown in Figure 6.3 (Chapter 6). We used the Jibo robot [168], which stands 11 inches tall and features three full-revolute axes for 360-degree movement. Jibo's onboard capabilities enabled naturalistic gaze and movement. A compact PC communicated with other hardware,

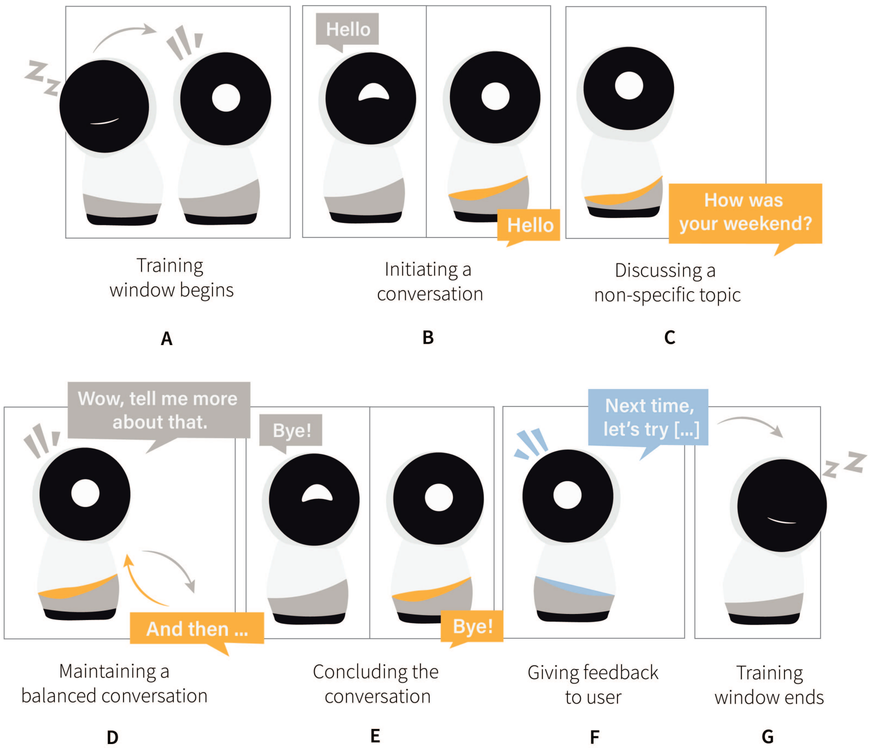


Figure 8.1: Training Sequence: A training session unfolds in distinct stages: (A) The robot "wakes up" to signal the start of the session, (B) encouraging users to initiate with a greeting. Users practice small talk skills, including (C) discussing non-specific topics, (D) maintaining balanced dialogues, and appropriate transitions (E). Users complete several conversations, with the robot giving micro-level feedback after each conversation (B-F). At the end of session, the robot gives the user macro-level, overall feedback (F). Outside of the specified training window, the robot performs an idling behavior (G).

monitored the system, and served as local data storage during deployment. An Azure Kinect camera was mounted two inches above Jibo's head to maximize the visual field and audio capture during training.

The system was designed for self-reliance, equipped with a mobile router that provides a prepaid internet connection for continuous WiFi and automatic cloud-based data synchronization. An uninterruptible power supply served as the charging station. With these components, the setup was plug-and-play, requiring only the charging station to be plugged in. To enhance ergonomic and accessible design, non-interfaceable components were encased in the container on which the robot was placed, reducing apparent complexity.

8.6.2 Software

We used a modular software architecture when creating the system to allow for individual components to be evaluated and improved. To achieve this modularity, we created the different components of our software as nodes in the ROS framework [433]. As illustrated in Figure 7.4, the system consisted of several components such as attention tracking of the users, robot behavioral control, and the generation of small talk dialogue. These components collectively contributed to small talk interactions described in our initial study (Section 8.4). In addition to these components, we introduced components of training presentation such as the session scheduler and mechanisms for generating relevant feedback to the user.

Delivering the Training Components. A scheduling node determined when the system would begin the training window. One or more training windows could be specified in the system's configurations. During the training window, the system would deliver a session, which consisted of engaging the user in several small talk conversations. At the end of each session, the system provided feedback, as detailed in Section 8.6.2.

Recent in-home robot deployments [3, 29, 34] have highlighted the importance of ensuring that in-home systems are capable of discerning when it is socially appropriate to engage users. These studies have emphasized that, for effective interaction, the system must recognize contextual cues that indicate whether the user is in a setting conducive to engagement or learning. Therefore, we incorporate two classifiers: audio-based social presence classification [34] and person detection. The Azure Kinect captured video recordings, and image snapshots were analyzed by a pre-trained YOLO neural network [480] to estimate the number of people present. The Kinect also transcribed audio using Google's Speech-to-Text API to identify speech content. If speech was detected, the social presence classifier determined whether it represented a physically co-present conversation (e.g., a dinner party or a conversation between friends) or a media interaction (e.g., watching television or listening to the radio) [34]. If two or more people were detected and the audio was classified as non-media, the system assumed it was not an appropriate time to engage the user and skipped the planned interaction. However, the frequency of planned conversations increased gradually to maintain a consistent number of interactions within the training window, with intervals selected from a Gaussian distribution to avoid predictability.

When the system was not engaging in a small talk conversation, Jibo performed an idling behavior, ranging from sleeping to looking down at the floor. When prompted to "wake up," Jibo looked up at the user to signal its availability to chat. The user could initiate conversation with a greeting, or Jibo would do so after some time. A linear-logarithmic growth function determined how long Jibo waited for user initiation, progressively extending this window. If the user does not respond, Jibo would prompt again.

Following a greeting, the system randomly selected 8 to 12 rounds of conversation. It used an observer model (Section 8.4) to ensure compliance with small talk criteria and monitor user progress. After the selected rounds are completed, Jibo transitioned to feedback mode, signaled by changing its LED ring to blue and adopting a more formal voice.

Generating Feedback. Feedback is essential in any training. In our earlier study [32], we found that quantitative definitions of small talk criteria (e.g., brevity, tone, non-specificity, coherence) effectively informed an observer model for generating feedback to the speaking model. However, these quantitative definitions do not easily translate into practical feedback for human users.

Effective user feedback should be specific, timely, constructive, objective, and respectful. We used a separate instance of GPT-3.5 that has a prompt delineating these characteristics of good micro- and macro-level feedback. *Micro-level feedback* highlighted user successes after each conversation, along with one area for improvement. For instance, "You asked great questions about my favorite hobbies. I would like to hear about your hobbies next time." *Macro-feedback* summarized all previous conversations within a training window, providing overarching insights. For example, "You did a great job acknowledging my interests; I enjoyed your rainy day activity suggestions. For our next conversations, try discussing your weekend plans or the weather." Additionally, if a small talk rule was consistently or severely broken during a training session, the robot would act out a dialogue to demonstrate following that rule. In these demonstrations, Jibo modulated its speech pitch and duration and changed its LED color. It even physically turned to "face" itself as it alternated between characters, like a puppet speaking to itself. Altogether, we aimed to explore if this generated feedback was effective, constructive, and well-received by adults with ASD.

Robustness for Contactless, Home Deployments. Robots deployed in homes require much greater robustness than those in controlled lab settings. The unstructured home environment poses challenges like power outages, fluctuating lighting, and unexpected user distractions. To enhance our system's reliability, we implemented watchdog scripts to monitor performance and remote desktop applications for troubleshooting. One script ran at the start of each training session to verify the camera, microphone, and communication with Jibo. The second script ran after each day to check the file sizes of recordings. The scripts would send an email detailing component success or failure. Remote access allowed for remote configuration; the system could be delivered to the user's home and then configured completely without in-person contact.

8.7 In-Home Deployments

The study was preregistered and received Institutional Review Board approval. Interested adults with ASD enrolled through the study's website promoted via channels that required a clinical diagnosis of ASD, such as ASD-specific residential facilities or employment networks. To comply with COVID-19 safety protocols, the system was designed for easy, contactless setup. It was delivered directly to participants' homes with detailed written and video instructions. In each home setting, informed consent was obtained from all participants and household members, with verbal assent obtained from minors. Remote support was available via phone or video call, and at no point did a researcher enter participants' homes.

Participants placed the system in a comfortable room and specified a daily training window of up to three hours. The study lasted at least seven days, though adjustments were made for scheduling conflicts. After the study, participants took part in an interview to discuss their experience with the system, its effectiveness, and suggestions for improvement.

8.7.1 Data Collection

Video and audio recordings of all training interactions captured participants' responses to the robot. It is well-known that automated systems, particularly those relying on computer vision or speech recognition, often face challenges with environmental noise or the diverse behaviors exhibited by participants, especially those with ASD [35]. As a result, we opted to manually annotate the collected dataset and verify the accuracy of inputs, such as speech transcripts, before applying automated measures. Using ELAN [484], three undergraduate research assistants annotated the

Figure 8.2: In-Home Deployments. The collage on the left shows in-home interactions with adults with ASD from the system's point-of-view. On the right is the system placed in a user's living room after the contactless delivery.

start and stop times for each robot and user response. These markers allowed us to calculate users' *initiation rate* defined as the proportion of conversations initiated by the user to all conversations within a session, and the duration of each *conversational turn* which is the time one speaker (robot or user) spent talking before the other responded. Additionally, for each turn, a binary label indicated whether the user made eye contact with the robot.

To ensure the reliability of these annotations, inter-rater reliability scores were calculated for each of the annotated metrics. Cohen's Kappa (κ) was used for categorical variables, such as eye contact, and the intraclass correlation coefficient (ICC) for continuous variables, such as initiation rate and conversational turn duration. The resulting κ for eye contact was 0.95, indicating a high level of agreement between the annotators in identifying whether the user made eye contact with the robot during a given conversational turn. For the continuous metrics, the average ICC for initiation rate was 0.91, and for conversational turn duration, it was 0.93. These scores demonstrate strong consistency among annotators in evaluating the training interactions.

Furthermore, transcripts generated by Google's Speech-to-Text were manually reviewed and corrected, particularly for users with accents or atypical intonations. Each response was assessed using the observer's evaluative metrics for small talk—brevity, tone, specificity, and coherence—as defined in Chapter 7.

These annotations and observer-derived metrics enabled us explore behavioral trends over the course of the study, focusing on the frequency of user eye contact, initiative, small talk violations, and overall conversational dynamics. This quantitative analysis was further enriched by qualitative insights from post-study interviews, which explored participants' comfort with the robot, their perception of their own conversational skills, and the relevance of the robot's training and feedback.

8.7.2 Participant Information

Twenty five adults with ASD (19 males, 6 females), ranging from 18 to 68 years (M = 32.4, SD = 12.7) participated in the study. All had a confirmed diagnosis of ASD and completed the Autism-Spectrum Quotient-10 (AQ-10) survey prior to the study. Fifteen participants were classified as high-functioning, with an average AQ-10 score of 4.8 ($SD = 1.2$). The remaining ten had higher scores ($M = 7.1$, $SD = 2.4$) and lived with caregivers. In this group, four participants were also diagnosed with Intellectual Disability (ID), four with Attention Deficit Hyperactivity Disorder (ADHD), two with Obsessive-Compulsive Disorder (OCD), and one with Down syndrome. All with co-occurring conditions were receiving medication and specialized therapy or education at the time of this study.

8.7.3 Results

A total of 2,114 conversations were recorded, yielding 9,870 user responses across 225 sessions or 281.3 hours of participant interaction. Participants experienced an average of 8.9 sessions (SD = 2.3) over the course of the study.

Initiating Conversations. We analyzed the frequency of user-initiated conversations across all sessions to understand the degree of proactive engagement with the system. On average, participants initiated conversations in 34% of interactions during their first session, rising to 64% by the final session. Users with higher AQ-10 scores started with a lower initiation rate of 14% in the initial session, which increased to 55% by the end of their study (t = 18.9, $p \leq 0.0001$). Conversely, high-functioning users showed an initiation rate of 50% in the initial session, which increased to 70% by the study's conclusion (t = 12.4, $p \leq 0.0001$). All 25 participants demonstrated some improvement in their initiation rate between the initial and final sessions of the study (ΔM = 36%, ΔSD = 20%).

A linear regression on the initiation rate over sessions yielded a significant increase ($\beta = 0.07$, $p \leq 0.0001$). This growth suggests that participants, irrespective of their initial engagement level, became more proactive in initiating small talk.

Maintaining Eye Contact. We assessed the effect of the training on users' ability to maintain appropriate eye contact. A linear regression analysis predicting eye contact frequency per conversational turn based on session number revealed a significant increase in participants' eye contact with the system over the course of the study ($\beta = 0.02$, $p \leq 0.001$). On average, participants maintained eye contact in 24% of turns during the first session, compared to 51% in the final session. Additionally, all 25 participants showed improvement in eye contact frequency from the initial to final session, with an average increase of 32% ($\Delta SD = 8\%$).

Sustaining Appropriate Small Talk. We chose to separate the greeting phase from the rest of the conversation because greetings, such as "hello" or "good morning," tend to be routine, shorter and less variable. In contrast, post-greeting responses are more varied, context-dependent, and reflect the true conversational dynamics. Greetings were identified through keyword matching and checking if the sequential response index is within the first full conversational turn. Post-greeting user responses (N = 6,751) were assessed using the observer's metrics and behavioral annotations.

Brevity. The average turn duration of user responses in seconds (s) exhibited a significant increase by session ($\beta = 0.77$, $p \leq 0.0001$). The average turn duration was 4.4s ($SD = 1.5$) in the first session as compared to an average duration of 5.3s ($SD = 2.8$) in their final session. This is supplemented by the observer brevity metric based on word count; the observer flagged 2,421 user responses (21%) as being overly verbose, and produced macro-level feedback on keeping responses more concise 61 times, or 27% of all sessions.

Further analysis revealed a significant increase in the robot's turn duration ($\beta = 0.79$, $p \leq 0.0001$), indicating that its responses lengthened over time. Users' conversational balance, defined as the proportion of turn duration between their previous utterance and current response, also improved significantly ($\beta = 0.10$, $p \leq 0.0001$). This suggests that users provided longer responses and engaged in more balanced, reciprocal interactions as the study progressed.

Non-specificity. The observer-derived metrics for non-specificity evaluate the frequency of named entities and descriptive words in interactions. Due to the nature of ASD and the associated challenges with theory of mind, it is common for individuals with ASD to mention named entities (e.g., a cat named Cheddar or a street called Cedar) without considering that the robot lacks the contextual knowledge needed to respond appropriately. In the post-study interview, the mother of P_{19} recounted, "Kurt often got confused, sometimes frustrated, when the robot misunderstood him.

For example, when he mentioned feeding his cat, Cheddar, the robot started asking about cheeses." Eight other users, all with high AQ-10 scores, demonstrated similar difficulties during their interactions.

A linear regression on named entity counts showed a significant increase by session ($\beta = 0.01$, $p = 0.002$), as did descriptive words ($\beta = 0.03$, $p \leq 0.0001$). Notably, only 16 participants exhibited these increases, with all but one categorized as high-functioning based on AQ-10 scores. The observer flagged 1,223 user responses (12%) as overly specific, providing macro-level feedback on broadening responses in 83 sessions (37% of total). Importantly, 85% of these violations occurred after the first three sessions, indicating that users, particularly higher-functioning adults, felt more comfortable engaging in deeper conversations as the study progressed.

Tone. There was no significant change in tone during the study ($\beta = 0.00$, $p = 0.69$). The observer metric for tone based on sentiment analysis yielded macro-level feedback on keeping responses more positive 27 times (12% of all sessions).

Coherence. There was no significant change in user's response coherence ($\beta = 0.00$, $p = 0.82$). The observer's coherence metric based on information gain yielded feedback on keeping responses relevant 54 times (24% of all sessions).

System Performance. A key indicator of the system's performance is sustained user engagement throughout the study. Above, we detailed trends in the frequency of user-initiated interactions and levels of eye contact with the robot. These are simultaneously considered measures of system performance and user improvement as a result of robot-assisted training.

Moreover, while our prior work in Chapter 7 showcased the effectiveness of the grounded observer in novel, in-person user interactions, we examine its performance in this context of in-home, long-term small talk training for individuals with ASD. Of all robot-delivered responses to the user ($N = 9,870$), the observer flagged $8,585$ instances in which the base model's output violated one or more overlay rules. This high rate of intervention[3] underscores the necessity of the observer framework; without it, a substantial portion of the LLM-generated responses would have been delivered to users despite failing to meet the requirements for appropriate small talk.

Analyzing the regeneration sequence provides insight into both the model's baseline behavior and the effectiveness of the observer's guidance. The first regeneration

[3]In our system design, we limited the feedback loop to a maximum of three regenerations to ensure the robot delivered a timely response. The number reported here reflects the cumulative number of first, second, and third retries through the observer's feedback loop.

reflects the base LLM's natural tendencies in the absence of any observer feedback. The second regeneration reveals how well the observer's initial feedback corrected the rule violation. Lastly, the third highlights persistent issues either reintroduced by the base LLM or not fully resolved by the observer's prior feedback.

A substantial portion of responses ($N = 6,415$) required at least one regeneration, indicating that the base model's initial outputs often failed to meet the overlay criteria. The most frequent reasons for rejection were excessive length (45%) and specificity (19%), which reflected expected tendencies of LLMs to produce verbose or overly specific replies. A second regeneration was required in 21% of cases ($N = 2,072$), most often due to issues with coherence (11%) and excessive brevity (6%). This suggests that the observer's initial feedback may have overcorrected the model's output, producing responses that were either disjointed or too terse. Finally, only 1% of responses ($N = 98$) required a third regeneration, with brevity emerging as the sole recurring issue. This indicates that while the observer's feedback generally guided the model toward acceptable outputs within just two attempts, a small subset of responses continued to fail its overlay rule for being verbose.

Though these "failed" responses reflect the efficacy of the observer in enforcing its overlay rules, this outcome is unsurprising for two reasons: (1) the brevity overlay was configured with relatively low rigidity, allowing flexibility in enforcement; and (2) as the deployment progressed, users tended to offer increasingly longer responses and were most often prompted to shorten their own replies. The system likely tolerated generating more verbose responses (as allowed by the less rigid overlay) to maintain relevance and alignment with user input. Additionally, the thresholds for identifying overly terse or verbose responses were informed by heuristics derived from interaction datasets gathered in Chapter 7, and thus remained somewhat arbitrarily defined. It is important to note that the observer was able to resolve all other rule violations within just two regeneration attempts, indicating effective feedback and compliance for rules beyond brevity.

Since the efficacy of system prompts is often assessed through post-hoc performance or trial-and-error testing, it is challenging to quantify the exact improvement that can be achieved through a more crafted initial or feedback prompts. Nevertheless, these results demonstrate that the observer was largely successful in realigning the base LLM's deviations from appropriate small talk conventions.

Overall, these results show clear improvements in users' conversational skills and engagement with the training system. As training progressed, users showed greater confidence and ease in initiating conversations, maintaining eye contact, and engaging in more balanced, yet detailed conversations over time. Interestingly, while longer, more descriptive responses were flagged as violations of small talk norms—resulting in the robot giving more frequent feedback to maintain brevity—these deviations suggest an increasing preference for more substantial conversations over time. Together, these results reflect users' growing sense of ease and interest in the robot as a conversational partner.

8.7.4 Post-Study Interviews

Here, we draw on insights from post-study interviews to contextualize the previously discussed quantitative results from participants' interactions with the robot. Interviews were conducted with 24 participants with ASD[4] and 12 primary caregivers, 10 of whom lived with a participant during the study. These semi-structured post-study interviews, conducted by a member of the research team, lasted between 30 and 45 minutes. The interviews were recorded and transcribed to ensure accuracy and participant confidentiality. Caregivers were invited to join the interview only with the prior consent of the participants. Based on the feedback collected, we performed an informal thematic analysis, categorizing the insights from participants and caregivers into three main themes, as outlined below.

Engagement with Training. Participants widely reported that the robot's structured yet dynamic interaction style helped make small talk practice approachable. Several participants mentioned that its unpredictable behavior (e.g., random sleep intervals and spontaneous prompts) mimicked real-life conversations, which kept interactions "fresh and challenging" (P_3). This aligns with previous findings that highlight the importance of unpredictability in training, as it more closely simulates natural, unscripted social encounters [241, 561].

Reviewing the transcripts, conversations broadly adhered to small talk topics (84% of all user responses), such as the weather and plans for the day. Topics beyond the scope of small talk included emotional conversations about experiencing bullying at work (P_{10}) or the recent passing of a family member (P_1), and technical interests such as steps to rebuilding a car transmission (P_3) or solving a homework assignment (P_{21}). While some users ($N = 9$) felt they had "quickly exhausted the limited range

[4]One individual was unable to participate in the post-study interview due to unforeseen personal circumstances.

of topics considered to be small talk" (P_3), others (N = 12) appreciated that even simple prompts about their day were "meditative, and made me reflect on the day and my feelings in a more mindful way," (P_2). Several users noted that, beyond skills training, the small talk robot provided valuable mental and emotional support. P_{20} shared, "My daily life can be isolating, so I don't often get asked how my day is going or how I'm doing [...] I enjoyed these kinds of chats when I'm making breakfast or coming home from a long day, watching TV, and there's no one else to talk to."

The robot also served as a valuable medium for facilitating communication between caregivers and their child. Several parents (N = 5) noted that the robot's "neutral, non-judgmental presence" (P_{12}) made it easier to initiate conversations on difficult topics. The parent of P_{14} shared, "We've had trouble talking about certain feelings, but when the robot asked about how his day was going, it opened up a way for us to talk. It felt like a safe space for him to express what he normally wouldn't share with me directly."

Perceived Skill Improvement. Adults with ASD (N = 18) and their caregivers (N = 10) remarked on behavioral gains which were not captured in the observer's metrics. The robot's training was designed to simulate real conversations, complete with pauses, turns, and varied topics, which helped participants grow more confident and adept at holding discussions. For example, P_{11}'s parent reflected, "My son has made great progress in understanding conversational turn-taking. He used to dominate conversations or struggle with waiting for his turn, but the training has helped him better grasp the flow of dialogue." Another parent observed how P_{17} became more confident in her interactions with others: "It's been incredible to see my daughter take more initiative in our family interactions. She's been practicing with Jibo, and now she actively participates in conversations at the dinner table, offering her own thoughts and even asking others about their experiences."

Relevance of Robot Feedback. Although many robot-assisted social skills interventions for individuals with ASD exist [21], most are designed primarily for children [59] and focus on practice rather than direct feedback. This is largely because social skills are personal and highly individualized [581], making it difficult to provide constructive feedback that resonates with each user. However, giving direct feedback is a core component of our training pedagogy. This approach to feedback could have broader implications for how we design robot-directed interventions for adult populations—while children may benefit from simpler forms of reinforcement, the robot's ability to provide explicit, personalized feedback was seen as highly valuable by many of our participants ($N = 21$).

This approach not only supported skills improvement but also encouraged cognizance of how one interacts with others. P_{10} shared, "During one session, Jibo and I discussed our favorite foods. The conversation lasted longer than usual because Jibo encouraged me to ask follow-up questions. By the end, Jibo gave me tips and even an example on how to keep conversations engaging—advice I've since used with my coworkers." Many users ($N = 13$) commented on being more mindful and reflective about the quality of their conversations as a result of the robot's feedback. P_{14} explained, "It was new to me, practicing this kind of mindfulness—thinking more deeply about how and what to say."

8.8 Discussion

This study presents an in-home, week-long deployment of a robot designed to support small talk training for adults with ASD. While most of the literature on robots for ASD intervention has focused on children, relatively little is known about the social needs of adults with ASD or how to support them effectively. Through this deployment, we address this gap by proposing small talk as a structured but flexible domain for practicing a range of conversational skills relevant to adult life.

To develop relevant training in which the robot can both model appropriate behavior and respond meaningfully to users, we first explored how to build robots capable of small talk interactions. To this end, we leveraged our grounded observer to allow the system to generate responses in real-time while enforcing key conversational norms. Enabling the robot to regulate its own social behavior was a necessary prerequisite for unsupervised, autonomous interaction with users in their homes over multiple days.

The small talk training robot was deployed into the homes of 25 adults with ASD for at least a week. Results indicate measurable improvements in users' conversational skills and sustained engagement with the system over time. As training progressed, participants demonstrated increased confidence and fluency in initiating conversations, maintaining eye contact, and producing more balanced yet expressive responses. In addition to these interaction-level findings, post-study interviews with participants and their caregivers provided insight into the real-world value of the system. Participants described feeling more comfortable practicing social exchanges with the robot and reported greater willingness to engage in similar situations with people.

Overall, this study contributes a novel, autonomous SAR system for adults with ASD and demonstrates how small talk can be leveraged as a practical and meaningful domain for building foundational social skills. Below, we discuss the ethical considerations that arose during our system development, limitations of the present study, and opportunities for future work.

8.8.1 Ethical Considerations

In general, the deployment of LLM-based systems in personal settings for vulnerable users raises ethical concerns. For example, as motivation for developing the grounded observer framework, our prior work described how inaccurate or misaligned responses could pose safety-critical risks (Chapter 7). Furthermore, the system must prioritize user privacy and security to prevent misuse of sensitive personal data.

At the time of this study, local use of GPT-3.5 and many other LLMs were not supported. While open-source alternatives offered similar architectures, they required extensive computational resources, making them impractical for real-time interactions that rely only on the robot's on-board capabilities and a compact PC. Other stable models, including GPT-2 [582] and LLAMA 2 [518], were trained on smaller datasets, resulting in less mature natural language capabilities and an increased risk of generating harmful language.

To further mitigate these risks, data was transmitted to the cloud only during active participant engagement in training sessions, minimizing the amount of participant data sent throughout the in-home deployment. While our informed consent process provided participants with a clear and accessible explanation of the cloud-based LLM's use, we encourage researchers and developers to thoughtfully balance a model's readiness for real-time interaction with considerations of privacy, security, and computational feasibility.

8.8.2 Study Limitations & Directions for Future Work

It is important to note that the present study was conceived as an exploratory investigation rather than a hypothesis-driven evaluation. Our primary objective was to demonstrate the *feasibility* and process of designing, developing and deploying a social robot platform for long-term, in-home interactions with an understudied population. Still, lasting behavioral changes from training typically require months or years to manifest. Although our approximately week-long intervention captured early engagement patterns and potential novelty effects, longer-term studies are needed to assess the sustainability of any observed improvements.

Further, given the practical challenges of establishing a comparable control condition in naturalistic settings, we adopted a within-subject design in which each participant served as their own baseline—a common approach in ASD research due to high inter-individual variability. We acknowledge the limitations of this method and refrain from making strong claims of efficacy. For instance, one may argue that increases in social initiations and eye contact could reflect familiarity or novelty effects. Interestingly, one may even expect the opposite—frequently repeated interactions, especially with a system of limited personalization or interaction memory, would result in *decreased* engagement over time. While we cannot attribute user outcomes solely to the system's design, the system shows promising potential in facilitating *continued* engagement with the robot-led training. Nonetheless, larger-scale studies with extended training periods will be essential for assessing the *efficacy* and long-term impact of this approach.

Moreover, this study was conducted within a single country and cultural context, which may limit the generalizability of its findings to other regions and populations. Differences in communication styles, social norms, approaches to the diagnosis or treatment of ASD, and attitudes toward technology across cultures can significantly influence both the acceptance of social robots and the effectiveness of the intervention [583]. To gain a more comprehensive understanding of the feasibility and efficacy of robot-assisted training, future research should replicate and expand upon this work in diverse cultural and geographic settings, accounting for the unique ways these factors may shape user engagement and outcomes.

Another limitation is that the behavioral detection and response generation of our robot system rely on models like Google's Speech-to-Text for speech processing and OpenFace for gaze and pose estimation, which are largely trained on data from neurotypical individuals in controlled environments. Consequently, deploying such models in in-home settings with adults with ASD introduces challenges due to the unique characteristics of this population and the unpredictable nature of real-world environments [35]. For instance, some participants in our study exhibited atypical speech disfluencies, such as irregular pauses, repetition, or unexpected intonation patterns, which are common in individuals with ASD [584]. These patterns can cause the system to misinterpret or entirely fail to recognize user inputs, leading to breakdowns in conversational flow. Similarly, inaccuracies in gaze estimation can result in the system failing to distinguish between the user and other visual stimuli in the environment, such as faces appearing on a TV or reflections in mirrors. This occasionally caused the robot to direct its gaze away from the user, potentially reducing the perceived quality and engagement of the interaction. Although these issues were not mentioned in our post-study interviews, we observed them while reviewing the dataset to correct errors in the generated speech transcripts. Given the growing interest in automatic behavioral annotation [585], our observations underscore the need for model adaptations or fine-tuning to better align with the needs of diverse populations and naturalistic settings.

8.9 Summary

Engaging in small talk is a vital social skill that impacts everyday interactions, from making friends to making good first impressions during a job interview. For adults with ASD, small talk poses unique challenges yet is essential for social integration and professional opportunities. In this chapter, we detail our development of an autonomous, in-home robot-assisted training platform to enhance small talk skills. While significant, long-lasting behavioral change typically require weeks or months of training, our results showed that even within a week, adults with ASD made meaningful progress, showing increased initiative in conversations and improved eye contact.

This study makes several contributions to our understanding of robot-directed social interventions. First, it demonstrates how relevant training can be designed to target desired skills for an understudied population. Second, it establishes the feasibility of creating a fully autonomous, plug-and-play, social robot for daily in-home practice. Third, it presents a system that sustains engagement with users over time by leveraging LLMs to generate context-aware, varied behavior; the results defy the expected decline in users' engagement typically observed as novelty effects wear off. Fourth, it highlights the importance of moving beyond rote rehearsal by incorporating personalized, directive feedback that mirrors real-world social exposure and learning. Finally, we underscore the importance of small talk—a form of communication often overlooked in therapeutic contexts—not only as a vehicle for social skills development, but also as a means of fostering positive mental, emotional, and interpersonal outcomes for users.

Now, we build on the technical work that enable these contributions—such as applying the grounded observer mechanism to constrain generative robot behavior for safe and effective deployment, disambiguating social presence in real-world contexts

[34], and delivering social feedback in contextually appropriate and naturalistic ways. The following chapter presents our design, development, and deployment of another robot-assisted intervention—this time aimed at facilitating emotional de-escalation among children in a public school environment. Aligned with the broader goals of this dissertation, this next chapter introduces a robot that targets novel regulation skills in a new, socially complex setting characterized by new real-world, ethical, and logistical constraints.

Chapter 9

From Fidgeting to Focused: Developing Robot-Enhanced Social-Emotional Therapy for School De-Escalation Rooms

The previous chapter presented a robot-assisted approach to support small talk training for adults with ASD. While large language models (LLMs) offer significant value in generating relevant training content, their integration into a robot introduces important technical and ethical challenges. For instance, how can we safely deploy LLM-driven systems to interact *autonomously* with *vulnerable users* in their *homes* for *extended* periods of time? To address these concerns, we applied the grounded observer framework (introduced in Chapter 7) to enforce behavioral constraints and ensure safe, context-appropriate interactions during deployment.

In this chapter,[1] we extend this line of inquiry to a new domain: supporting regulation among multiple users with diverse cognitive profiles in a public setting. Specifically, we focus on emotional de-escalation within a public elementary school. Many schools have built de-escalation and sensory rooms to support students who experience heightened emotional states, sensory overload, or difficulty self-regulating in traditional classroom settings. Yet, effective implementation remains challenging due to diverse student needs and resource constraints. To address this gap, we developed a robot designed to facilitate self-regulation within a school's existing de-escalation space.

We now move from the home setting of our previous deployments, where the costs of dysregulation may be more diffuse or private, into a more public setting where dysregulation presents higher-stakes risks across academic, physical, and social domains.

[1]This chapter is adapted from our published work: **Ramnauth, R.**, Bržčić, D., & Scassellati, B. (2025, August). From Fidgeting to Focused: Developing Robot-Enhanced Social-Emotional Therapy (RESET) for School De-Escalation Rooms. In the *34th IEEE International Conference on Robot and Human Interactive Communication* (RO-MAN). IEEE. [31].

These factors introduce new design challenges: the robot must respond flexibly to a wide range of user needs, operate without assumptions about age or diagnosis, engage both new and returning users, and uphold the social and institutional norms of the school environment.

This chapter presents our co-design process, iterative development, and final system architecture. Following a fully autonomous, month-long deployment in an elementary school, we assessed the robot's usability and impacts. Results indicate our robot integrated well into the school environment, promoting more efficient de-escalation, smoother transitions back to classroom learning, and lasting impacts beyond its deployment period.

9.1 Introduction

Abby,[2] a bright seven-year-old with a passion for dinosaurs and all things prehistoric, entered second grade at her new school in Brooklyn, New York. From the first day, her behavior stood out from her peers. While she could focus intently on independent work related to her interests, she struggled to participate in group tasks, often covering her ears and withdrawing. The bright posters, colorful decor, and constant movement, which were appropriately stimulating for many children, overwhelmed and distracted Abby.

Her teachers misunderstood her withdrawal as defiance or disinterest, while her meltdowns were seen as tantrums rather than reactions to sensory overload. Despite the school's attempts to support her, such as assigning a paraprofessional to remove her from the classroom during stressful moments, Abby continued to struggle. Her overstimulated state made it hard for her to express her needs and feelings, limiting the effectiveness of those interventions.

Eventually, Abby's teachers recommended that she transition to District 75, the specialized district for students with significant disabilities and special needs. This decision was intended to create a more manageable classroom environment and ensure Abby received the individualized support she needed. Meanwhile, Abby's parents pursued a clinical diagnosis, confirming her Autism Spectrum Disorder (ASD).

Schools across the country have seen many students like Abby. When she started second grade in 2014, the rate of developmental disabilities in the U.S. was 1 in 17 children [586]; today, it is 1 in 6 [587]. To this, public settings like schools,

[2]Based on the personal account of the parents to the authors. Throughout this chapter, student names are anonymized and school omitted for privacy.

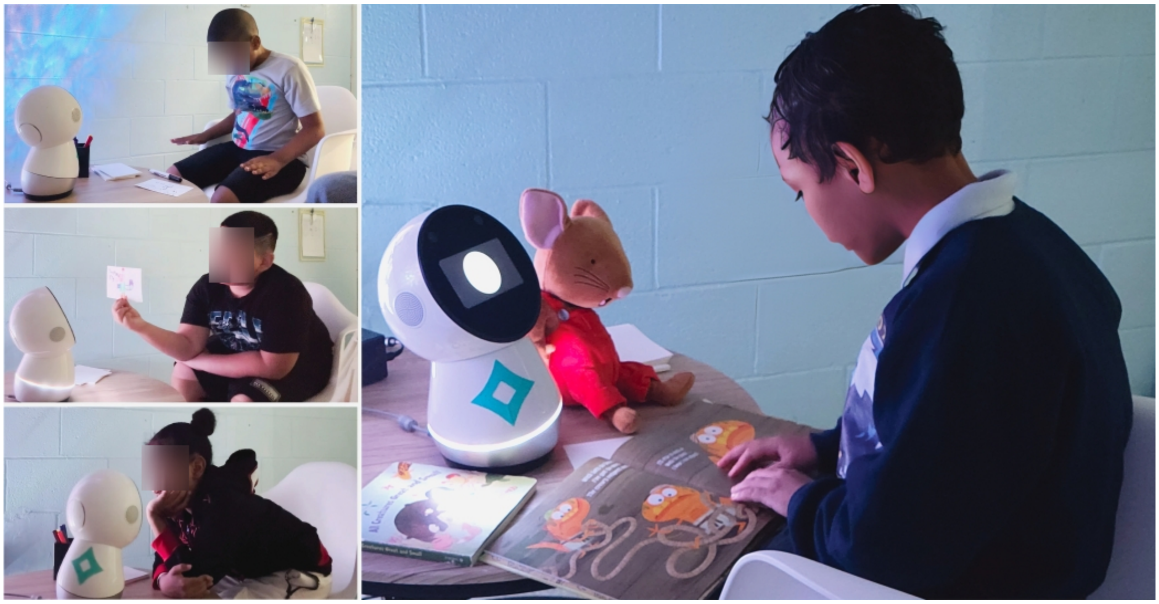


Figure 9.1: Robot Interactions in Schools. Students organically interacted with the RESET robot in their school's de-escalation room, engaging in activities such as guided deep-breathing exercises, small talk, and collaborative storytelling.

hospitals [588], and shopping malls are increasingly adopting sensory or *de-escalation* spaces to support individuals with social-emotional or sensory processing needs [589]. Equipped with tools like weighted blankets and tactile manipulatives, these spaces are designed to be calming environments that promote self-regulation, reduce anxiety, and enhance social participation. This approach aligns with broader inclusive efforts, such as integrated co-teaching (ICT) [590], where those with disabilities can function and learn alongside typically developing peers.

However, implementing de-escalation spaces presents several challenges and considerations that must be carefully managed [591]. For example, these spaces require significant resources such as specialized materials, available physical space, and additional training and staff. Also, these spaces must be carefully crafted to facilitate diverse goals—from reducing anxiety to supporting core socio-communicative skills like self-regulation—without introducing additional overstimulation or distraction [592]. Another concern is inadvertently reinforcing negative behaviors if users perceive the space as a reward for misbehavior rather than an intervention [591]. Thoughtful planning and ongoing evaluation are essential for these spaces serve their intended purpose effectively.

Socially assistive robots (SARs) have the potential to enhance educational and remedial efforts by providing personalized, physically co-present interactions [454].

SARs that are even sensorily minimalistic have proven to foster core social and emotional competencies [29, 35], enhance learning outcomes [593], and reduce stress [594]. These systems represent consistent, non-judgmental social partners that can adapt in-situ to users' diverse sensory and social needs.

Hence, we developed a robot-assisted intervention, RESET (Robot-Enhanced Social-Emotional Therapy), to engage with students in a school de-escalation room. We assess its feasibility and effectiveness in fostering students' self-regulation skills while furthering the room's intended goals.

9.2 Background

In 2022, the United Federation of Teachers and U.S. Department of Education (DOE) launched the "Sensory Tools for Healing Schools" program, distributing kits—consisting of items such as fidget toys, beanbag chairs, and noise-blocking headphones—to over 16,200 public school classrooms [595]. These kits aimed to provide accessible, immediate interventions to calm students, allow them to self-regulate, and improve their engagement with learning. Building on this initiative, many schools have since established dedicated spaces for de-escalation and sensory regulation [592].

9.2.1 The Role of a De-escalation Space

While no formal guidelines exist for designing effective de-escalation spaces, the school psychologist at our deployment site summarizes their role: "Sometimes, the most appropriate intervention is to remove a student from a context, equip them with the necessary space or skills, and return them to the original context to apply themselves differently." To facilitate the ease in which support staff can conduct these interventions, her school repurposed a storage closet (Figure 9.2), incorporated soft lighting, sensory tiles, and positive affirmation posters. Also, academic activities like books were included to support continuity with typical classroom tasks.

Given the limited data on these spaces' impact, we surveyed 115 educators—including 51 teachers, 23 administrators, 4 school therapists, and 37 support staff—across four DOE elementary schools with dedicated de-escalation spaces during the 2024 academic year. From free-form responses, we thematically organized the key factors that contributed to their space's efficacy, such as "independent activities that easily capture students' attention" (P_{14}; 77% of all responses), "minimalistic stimuli to reduce distractions" (P_{81}; 56%), and structured activities for 1:1 social skills practice

Figure 9.2: Room Integration. Examples of spaces currently in DOE public schools. The first image shows a closet space that was eventually converted into the de-escalation space shown in the second image. Items such as beanbags, tents, and plush animals are integrated into the school's library (third), or in a classroom corner (fourth).

(95%), such as turn-taking, cooperation, and language development. Many described using the space to "refocus students' attention" (P_3; 87%), "as a proactive strategy to prevent behavioral issues" (P_{14}; 55%), and "facilitate integration" (P_{67}; 76%) without disturbing ongoing classroom learning.

A space's efficacy can be evaluated using three primary metrics: (1) *cooldown*, the time from a child's entry to their engagement in an activity (described by 81% of survey respondents); (2) the extent to which student goals—such as social skills practice or completing a targeted activity—are met (91%); and (3) the generalizability of behaviors observed in the space to other settings, such as the classroom (53%).

9.2.2 Cognitive Barriers to Self- or Assisted Regulation

Self-regulation is a core executive function that enables individuals to manage emotions, attention, and behavior in response to changing environments. While de-escalation spaces provide relief for students experiencing stress or heightened emotions, transitioning from this to focused attention is cognitively demanding. This skill develops over time, and young children—particularly those with ADHD, ASD, or trauma histories—often have weaker self-regulation mechanisms [596]. According to cognitive load theory, when working memory is overburdened, the brain struggles to retain new information or make intentional decisions [597]. Thus, even when students are physically removed from a stressful context, they may be unable to engage with self-regulation strategies unless those strategies are explicitly scaffolded to match their immediate cognitive state [589].

While reducing external stressors is essential for de-escalation, over-reliance on this strategy may introduce issues. For instance, if students learn that dysregulation leads to access to a preferred escape from classroom learning, they may escalate behaviors rather than develop internal regulation skills. Given these constraints, effective de-escalation interventions must balance two critical needs: (1) providing structured support for cognitive and emotional realignment, while (2) ensuring elements do not hinder reintegration into learning spaces. Research on sensory-informed intervention models [598] recommends integrating predictable, low-stimulation activities that reduce cognitive demand while promoting gradual self-regulation, such as guided breathing, simple tactile engagement, or structured conversations [592].

9.2.3 Potential Role of Social Robots

Robots exist on a spectrum of sociability, ranging from toy-like devices that elicit minimalistic cues to human-like entities capable of rich, multi-modal interaction. By modulating where a robot is on this spectrum, it can be designed to offer safe, predictable, and structured support whilst fostering core socio-emotional skills [20]. SARs have been proven to effectively mitigate stress [594], deliver cognitive behavioral therapy [599], enhance learning outcomes [368], and provide peer companionship [325]. Hence, robots have the potential to effectively support school de-escalation goals.

9.3 Co-Design of the Robot

Our study was conducted in collaboration with a New York City DOE public school, serving students from kindergarten through fifth grade (K-5), with an approximate enrollment of 600 students in the 2024 academic year. We adopted a participatory design approach involving the school principal, the guidance counselor, three assistant principals, the school psychologist, and four ICT classroom teachers. This partnership was structured around iterative co-design sessions and needs assessment workshops to ensure the design of the RESET robot was informed by the lived experiences of those directly supporting students' socio-emotional integration. We outline the seven identified design objectives below.

Realistic. The system should provide timely, context-aware responses to users, enabling unscripted, multi-turn conversations with naturalistic gaze and motion.

User-Driven. RESET should prioritize student agency, allowing users to initiate and control their engagement with the robot, rather than enforcing rigid interactions (c.f. [600,601]). Research shows that user control enhances engagement, learning, and participation [602].

Structured Tasks. RESET must deliver structured, time-limited interventions that support users while minimizing missed class time, guiding them through goal-directed, sequenced activities using known de-escalation strategies [603].

Repeatable. While interactions should remain structured, RESET must generate sufficient behavioral variability to sustain engagement over time. Given the room's frequent use, RESET's responses should be varied enough to prevent habituation while preserving consistency in core functions.

Easy Integration. RESET should smoothly integrate into the broader therapeutic ecosystem, complementing rather than replacing existing tools and strategies. Students should interact freely with the room, robot, or staff facilitators as needed. In a diverse K-5 setting, interaction content must also be relevant, accessible, and engaging for all age groups.

Embodied. The system should be embodied as a robot. Research proves embodied robots yield measurable learning gains [368], enhance compliance [474], and provide salient social cues that encourage appropriate user responses [35].

Autonomous. RESET must operate autonomously without requiring technical expertise to set up or maintain.

9.4 System Components

Here, we outline the system's hardware and software supporting our design goals. Before deployment, co-design partners tested the full interaction sequence at least twice, often with selected students. Over the course of a week and 24 sessions, we iteratively refined the following components.

9.4.1 Hardware

Our RESET system consisted of six primary hardware components. We used the Jibo robot [168], an 11-inch-tall device with three full-revolute axes for 360-degree movement, enabling naturalistic gaze and motion. Based on feedback during our formative testing, we later included a textured, four-point star sensory sticker on Jibo's body and a projector to facilitate interactions (Section 9.4.3). A compact PC enabled communication between hardware components, monitored system performance, and managed data storage during deployment. To enhance vision and audio capture quality, an Azure Kinect was mounted two inches above Jibo's head.

For self-sufficiency, the system included a mobile router with a prepaid internet connection for continuous WiFi and cloud-based data synchronization. An uninterruptible power supply served as the charging station, making the setup plug-and-play with a single power connection. Non-interfaceable components were encased and secured to the underside of the table upon which the robot was placed, minimizing visible complexity. For safety, the robot was securely fastened to the table to prevent students from picking it up. If relocation was necessary, an adult could unlock the fastener using a key.

9.4.2 Software

We used the ROS framework [433] to evaluate and refine individual system components, including presence detection, robot actuation, observer-enabled dialogue, and scheduling.

Presence Detection. Recent real-world deployments [29, 34] have underscored ensuring that systems can recognize when it is socially appropriate to engage users. These studies highlight that effective interaction requires interpreting contextual cues to determine whether the user is in a state conducive to engagement or learning. To achieve this, we implemented two classifiers: person detection and audio-based social presence detection [34]. A pre-trained YOLO neural network [480] analyzed the video capture of the Azure Kinect to estimate the number of people present. Additionally, audio transcribed using Google's Speech-to-Text API determined speech content. If speech was detected, the social presence classifier distinguished between a physically co-present conversation (e.g., between a child and teacher) and a media-based interaction (e.g., music played on a tablet) [34]. If three or more people were present and speech was non-media-based, the system assumed an ongoing social activity and remained idle; otherwise, it initiated interaction.

The motivation to use an external camera rather than Jibo's on-board cameras include the stability of a stationary capture and to enable presence detection even when the robot is not looking. For instance, we designed subtle motions for its idle state, such as "sleeping" or glancing down at the floor. Also, the video feed was processed using OpenFace [434] to extract users' gaze orientation. This gaze data was then used to trigger robot-initiated interactions upon eye contact or to determine when to switch contexts, as discussed later.

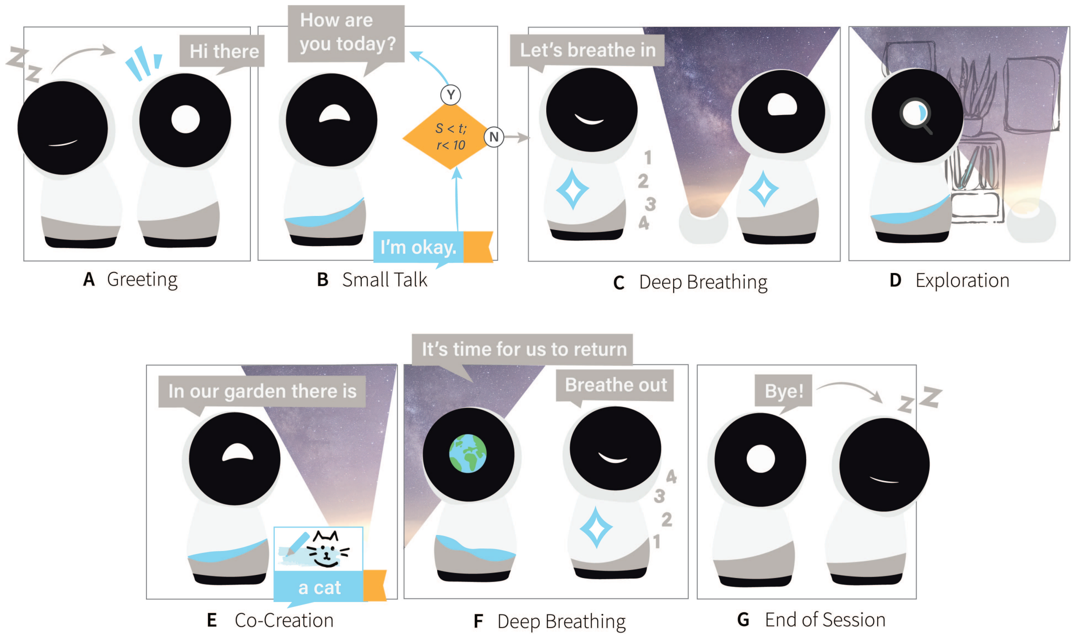


Figure 9.3: Interaction Timeline. The session unfolds in stages, starting with small talk and progressing through sequenced activities, including guided breathing exercises, storytelling, exploratory tasks to promote situational awareness, a co-creative drawing task, and concluding with a reflection on the interaction before transitioning students back to their classroom.

Dialogue Generation. To generate contextually appropriate dialogue in this domain, simply using an off-the-shelf large language model (LLM) to interact with vulnerable children is insufficient without ensuring behavioral safeguards. Thus, we build on the theoretical framework introduced in our prior work (Chapter 7), which involves applying overlay rules across the partially defined action space of a separate LLM instance, enabling an "observing" LLM to evaluate and correct the speaking model's behavioral deviations.

In line with our design goal of simulating realistic, multi-turn casual conversations, we replicate the small talk system described in our prior work, incorporating overlay rules that constrain brevity, tone, thematic coherence, and topic non-specificity. If the generated response meets these criteria, it is relayed; otherwise, the observer generates a revised system prompt and provides feedback to the speaking model. This feedback loop allows the system to self-correct when deviations are detected, whether from the LLM's forgetfulness of its initial system prompt or user-induced changes.

During our pilot testing, our co-design partners highlighted that older students (grades 3-5) tend to find multi-turn conversational interactions more effective for remediation, while younger students (K-2) are more likely to engage in parallel play—

playing near but not necessarily with others—and often use conversation as a secondary activity while engaging in peripheral tasks [604]. Additionally, younger children benefit from brief, dynamic interactions supported by visual and auditory cues, while older students are more suited to goal-oriented, structured conversations that provide clear feedback. This presents a design challenge: how to ensure the system's speech and interactional content are both relevant and accessible across a broad range of developmental stages.

To address this challenge, the observer's permissible limits for brevity (measured in words spoken) and coherence (calculated as the information gain between BERT-derived token embeddings [527] and that of the prior response) were dynamically adjusted according to a reinforced exponential decay function to mirror the user's speech. For example, if a kindergartener consistently engages with brief, disjointed replies, the robot progressively shortens its own replies and becomes more accepting of thematic incoherence between conversational turns, providing support without demanding rigid exchanges. Conversely, if a fifth-grader details difficulties about a group science project, the system can generate longer, more descriptive, yet thematically coherent responses.

Managing Context Switching. The ultimate objective of the system is to reduce missed classroom learning while remaining flexible enough to accommodate interventions based on students' cognitive learning goals or socio-emotional state. Hence, its overall interaction must be time-limited, yet appropriate to encourage students' smooth transitions back to the classroom as well as future use of the de-escalation room. Thus, given these constraints, the system should know how and when to appropriately switch contexts.

To manage the duration of a dialogue, the system initially selects a baseline of three to five rounds of small talk, which can be dynamically adjusted using a similar reinforced decay function based on the output of a fixation detector. Using the aforementioned overlay rules for topic specificity (determined by NLTK's named entity chunker and part-of-speech tagging [526]), tone, and coherence, the fixation detector first evaluates whether the user is maintaining strict continuity in the conversation. This continuity can either be productive, such as describing a classroom lesson, and require redirection, such as speaking aggressively or perseverating unhealthily on a single topic. The detector then differentiates between appropriate and inappropriate continuity using a keyword dictionary, calculating the semantic similarity based on BERT-derived token embeddings of the user's responses in relation to the dictionary.

From this outcome, the system can infer whether shifting the interaction to a related or entirely new topic would be beneficial.

To avoid forcing interactions, if the user does not respond to a communicative bid, RESET will issue a verbal follow-up. It would then provide up to two more non-verbal prompts before returning to idle. Users can also prompt shifts in dialogue or activity (detailed in Section 9.4.3), such as "I don't feel like talking today," processed by keyword matching.

Robustness for Deployment. Real-world settings demand greater system robustness than controlled lab settings. The dynamic and unstructured school environment poses challenges such as power outages, fluctuating lighting, noise, and diverse spontaneous interactions with children. Hence, we implemented remote troubleshooting and watchdog scripts. One script ran thrice daily to check hardware connectivity, while another verified data file sizes after each day. An automated status summary was sent via email. Remote access enabled the system to be delivered to the school, configured, and maintained all without disruption from researchers.

9.4.3 Interaction Design

The interaction was iteratively co-designed during the pilot phase to ensure the system can support a smooth progression through tasks that encourage both cognitive and emotional regulation as needed. While the sequence is highly structured to align with our design goals, it also incorporates system variability and adaptability to individual users at each stage.

Considering children's potential attachment, we also designed RESET's introduction and exit strategy. We incorporated a backstory into its dialogue generation, allowing RESET to introduce itself as visiting from another planet when greeting users or explicitly asked. This approach helped manage its presence during the study and aimed to make its eventual removal as minimally disruptive as possible.

A session begins when the system is woken from its idle state by users' presence (Section 9.4.2), Jibo's touch sensor, or its default wake command, "Hey Jibo." The robot then lifts its head to acknowledge the user (Figure 9.3-A). The user can initiate a dialogue, or RESET may do so at a random interval (Gaussian, 5 ± 2 seconds) to appear organic. The **greeting phase** includes a brief small talk chat (Figure 9.3-B).

After small talk, RESET initiates a box-method **deep-breathing exercise** (inhale, hold, exhale, each for four seconds; Figure 9.3-C). The robot encourages the user to place their finger on its textured star, close their eyes, and trace the star to help Jibo "travel" through space. During this exercise, Jibo displays a half-moon eye and gently moves up and down, while chimes cue the breathing rhythm. The projector gradually brightens to show a newly generated galaxy.

Imagining they've arrived on a new planet, RESET encourages the student to explore, such as by helping search for "robot friends" (Figure 9.3-D). While we did not provide props for this activity, we anticipated students would creatively use objects within the de-escalation room, allowing for a more flexible and personal experience without implying a "correct" way to respond. As the robot also models looking around the room, this exercise promotes **situational awareness**, inspired by the five-finger method for managing anxiety [605].

To add to their new planet, they collaborate on a drawing task (Figure 9.3-E). With open-ended prompts and encouragement from the robot, the student imagines and adds elements to their drawing. RESET may intermittently offer up to three minor suggestions to help maintain focus. Although there is no visual detection of the drawings due to potential inaccuracies given real-world constraints, RESET occasionally asks the child about the drawing, colors, and details so it can not only respond meaningfully but also promote **metacognitive reflection**, fostering a deeper sense of ownership in the **co-creative task**. After the time limit or once the child has described a set number of objects, RESET asks to "take a photo" of their creation, prompting the child to put down their materials and lift their artwork (Figure 9.1).

Finally, RESET cues the child that it is time to "return to earth." Together, they perform another breathing sequence, while the projector gradually turns off (Figure 9.3-F). RESET then celebrates the session's end by highlighting two positive moments with the user, reinforcing a sense of personal accomplishment before transitioning the child back to class (Figure 9.3-G). A complete session lasts up to 20 minutes total.

9.5 Deployment

During the 2024 academic year, we conducted a deployment of the RESET robot in the school's de-escalation room spanning a full instructional month (four weeks, or 20 class days).[3] The robot was placed on a designated table, as shown in the second image of Figure 9.2, where students could naturally encounter and interact with it. Existing room activities included fidget toys, plush animals, electronic tablets, and

[3]The study, including co-design sessions, iterative development, and deployment, was approved by both university and NYC DOE IRBs.

learning materials resembling typical classroom activities.

9.5.1 Data Collection

To assess RESET's impact, we combined system logs, existing school documentation practices, and post-hoc annotations and interviews to capture both quantitative and qualitative measures of student engagement and self-regulation.

Automated Logs. RESET autonomously logged interaction frequency, duration, and content. With appropriate consent, video and audio feeds were analyzed locally to examine speech patterns and gaze behavior during interactions.

Documentation Practices. We incorporated data from the school's established documentation practices for the room's use. These records included staff notes describing how their students engaged with the space, behavioral or classroom goals specifying the intended outcomes of their visit to the room, and sign-in and sign-out times. Students also completed self-assessment "temperature" charts, ranging from red (high distress) to green (calm and ready-to-engage), to indicate their readiness to use the space[4] or return to class.

Cooldown Annotations. The *cooldown period* is the time from when a student enters the space to when they exhibit observable signs of emotional and behavioral regulation, indicating readiness to return to the classroom (Section 9.2). It was assessed post-deployment via video annotations by trained staff. Annotators[5] identified this metric based on appropriate physical engagement with activities (e.g., holding a book or sitting with headphones). Final cooldown periods were calculated as the average of three independent annotations for each visit, with high inter-rater reliability ($ICC = 93.7\%$).

Pre- & Post-Deployment Phases. We conducted an ABA analysis, examining visits one month pre-deployment, during the robot intervention, and one month post-deployment to assess baselines, intervention effects, and lasting impacts after the system was removed. To complement quantitative interpretation, we conducted semi-structured interviews with teachers and support staff to gather insights on RESET's effectiveness, limitations, and areas for improvement.

[4]Per existing school protocol, if a student cannot complete their chart, such as due to a meltdown, the accompanying staff redirected them to safe, alternative activities outside the de-escalation space.

[5]The school's guidance counselor, psychologist, and assistant principal volunteered to timestamp this metric using the ELAN software [484].

9.5.2 Participant Information

A total of 57 students participated in the study, with many engaging with the robot multiple times during its deployment. By psychological age groups, 32 students (56%) were in early elementary (grades K–2), while 25 (44%) were in upper elementary (grades 3–5). Of the total, 26 students (46%) had formal diagnoses, including ASD and ADHD. Furthermore, 34 (60%) had an Individualized Education Program (IEP),[6] with 14 focused on sensory regulation, 7 on attention management, and 29 on social skills development.

Those with IEPs used the room more frequently, primarily by schedule, averaging 1.6 (± 0.4) visits per week. In contrast, the 23 students without IEPs averaged 2.5 (± 1.0) visits per month, often accessing the room through self-directed breaks or pre-arranged agreements with their teacher. Regardless of referral source, 93% of visits across the academic year indicated student stress, with minimal monthly variability ($\pm 5\%$)—excluding the deployment month and afterward, when the rate was $85 \pm 9\%$. During the deployment, nine staff, including four ICT teachers, two school therapists, and three paraprofessionals, accompanied students in the room.

9.5.3 Results: Visit, Interaction, and Cooldown Durations

Over the course of the deployment, a total of 295 visits were recorded in the de-escalation room, with 278 student interactions with the robot. Visits averaged 17.0 minutes (m; ± 7.7), and interaction time with robot averaged $7.8m$ (± 4.8).

Two Poisson generalized linear mixed-models examined the effect of week on visit and interaction durations, with a student identifier as a random effect to control for individual variability. The random effect was significant in both models ($\sigma^2 \geq 0.04$, $p < 0.001$), indicating substantial differences in visit and interaction durations among students. Visit duration remained stable across the first three weeks ($\beta \leq -0.09$, $p \geq 0.09$), with a small but significant increase from Week 3 to 4 ($\Delta M = 1.3m$, $\beta = 0.09$, $p = 0.04$), suggesting a marginal rebound in visit times in the last deployment week.

In contrast, robot interaction duration exhibited a clear week-to-week decline ($F = 43.9, p < 0.001$). Durations dropped from Week 1 to 2 ($\Delta M = -2.2m$, $\beta = -0.18$, $p = 0.02$) and further decreased into Week 3 ($\Delta M = -0.7m$, $\beta = -0.37$, $p < 0.001$) before stabilizing in Week 4 ($\beta = -0.11$, $p = 0.12$). This potentially suggests that

[6] A mandated plan to provide specialized instruction and accommodations for individuals with disabilities [606]. Students may have multiple IEP goals.

the deployment period was sufficiently long to capture both initial novelty-driven engagement and subsequent stabilization.

To minimize missed classroom time, it is critical that students efficiently achieve their goals in the de-escalation space and transition quickly back to classroom learning. A Kruskal-Wallis test revealed an effect of study phase on visit duration ($H = 298.0, p < 0.001$) indicating significant differences across all three phases, pre- ($N = 190$), during-, and post-deployment ($N = 220$). Pairwise comparisons showed that visit duration was significantly longer pre-deployment ($M = 31.6 \pm 7.4m$) but decreased significantly during the robot deployment ($M = 17.0 \pm 7.7m$, $p < 0.001$) and remained stable post-deployment ($p = 0.99$). These findings suggest that the robot's introduction led to a sustained reduction in visit duration, indicating a lasting shift in the space's use.

A goodness-of-fit revealed significant differences in visit frequency across phases ($\chi^2 = 17.5, p < 0.001$), with visits increasing during deployment. This suggests the robot's presence may have encouraged more frequent visits. However, the deployment month overlapped with the annual statewide testing period, a time of increased stress and classroom disruptions, which may have also impacted visit patterns.

A Kruskal-Wallis test revealed a significant effect of phase on cooldown period ($H = 284.7, p < 0.001$). Pairwise comparisons showed significantly shorter cooldown times during deployment ($M = 5.4m$) than both pre- ($M = 14.6m$, $p < 0.001$) and post-deployment ($M = 8.1m$, $p < 0.001$), though post-deployment times remained lower than pre-deployment. These findings suggest the robot's presence significantly reduced cooldown duration, with some lasting improvement in self-regulation efficiency even after removal.

9.5.4 Results: Documented Visit Goals and Activities

Existing documentation practices showed diverse visit goals, including sensory regulation (24%), attention management (12%), and social skill development (64%). These distributions remained uniform across all three phases ($\chi^2 = 0.2, p = 0.99$) indicating that the robot's presence did not significantly influence the reasons students visited the space.

Students participated in a range of activities across the three deployment phases, with a significant shift in activity distribution ($\chi^2 = 10.7, p = 0.002$). Activities were organized into three categories: *passive* regulation (e.g., fidget toys, quiet reflection), *active* (e.g., stretching, jumping), and *social* strategies (e.g., chatting, collaborating).

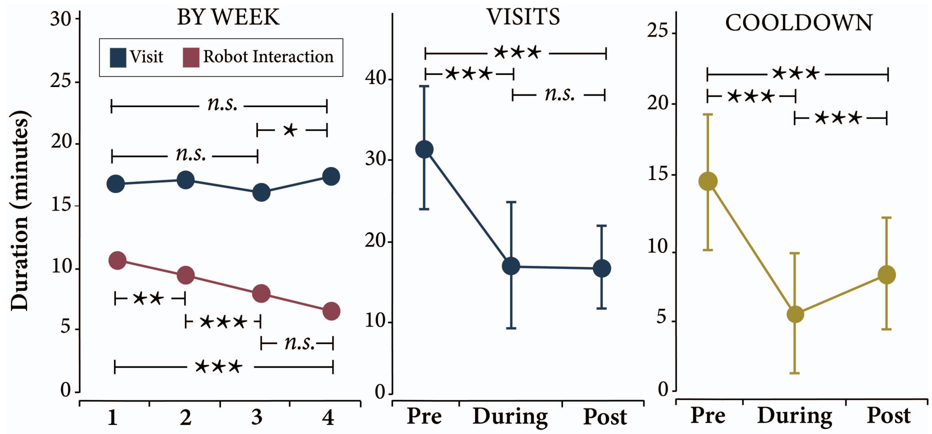


Figure 9.4: Deployment Outcomes. RESET led to shorter visits, faster cooldown and transitions back to class. The time needed for robot-assisted de-escalation decreased each week.

Pre-deployment, passive strategies were most common (62% of visits), followed by active regulation (45%). Social strategies were few (27%), as most students regulated independently.

During the deployment, social activities significantly increased (84%, $p = 0.005$), as students or staff opted for interactions with RESET. This shift led to a decrease in active regulation, as students engaged with more stationary, structured, social strategies. Passive activities also declined to 46%, although some visits incorporated robot prompts or mere presence. For instance, rather than silently completing assigned reading, students read aloud to RESET and discussed the material during its small talk (shown in Figure 9.1).

In the post-deployment month, passive strategies increased to 68% ($p = 0.001$), reflecting a return to more independent activities. Social activities declined to 43% ($p = 0.03$) but remained higher than pre-deployment levels, while active strategies remained lower (29%, $p = 0.01$). These results suggest the robot promoted social engagement as a primary regulation strategy, but following its removal, visits largely reverted to passive, non-social approaches. However, the sustained increase post-deployment suggests the robot may have fostered a longer-term shift toward social strategies.

9.5.5 Results: Educator and Staff Evaluations

Staff members ($N = 51$)[7] who supported students in the de-escalation room completed a survey of 12 items rated on a 5-point scale (1 = strongly disagree, 5 = strongly agree), assessing RESET in four key areas: student engagement, self-regulation support, classroom reintegration, and ease-of-use.

Overall, staff reported moderate to high agreement that the robot facilitated student engagement, with 74% agreeing or strongly agreeing that students appeared more engaged in self-regulation activities when the robot was present ($M = 4.1 \pm 0.9$). Similarly, 68% of staff felt that the robot helped students transition into calming activities more smoothly ($M = 3.9 \pm 1.0$), and 59% believed the robot made students more willing to use the de-escalation room ($M = 3.7 \pm 1.2$).

In terms of regulation efficacy, 66% of respondents agreed that the robot helped students regulate their emotions more quickly ($M = 3.8 \pm 1.0$), and 72% observed that students demonstrated improved self-regulation strategies while interacting with the robot ($M = 4.0 \pm 0.9$). However, responses were more mixed regarding its efficacy for those with sensory regulation needs, with 54% agreeing it was beneficial ($M = 3.5 \pm 2.8$), but 26% remaining neutral and 20% disagreeing.

When assessing classroom reintegration, 94% of staff agreed that students returned to class in a more regulated state after interacting with the robot ($M = 4.1 \pm 0.5$). However, only 47% felt that students carried over the strategies learned with the robot into their classroom behavior ($M = 3.4 \pm 1.1$). Notably, 42% reported that the impact of the robot on students' self-regulation remained noticeable even after the robot was removed ($M = 3.2 \pm 1.2$).

Regarding feasibility and effort, 70% of respondents found the robot easy to integrate into existing de-escalation practices ($M = 4.0 \pm 0.8$), and 62% agreed that it required minimal additional effort ($M = 3.7 \pm 1.0$). When asked about long-term use, 68% of staff members expressed support for continuing the robot's deployment in the de-escalation room ($M = 3.9 \pm 1.1$), though some noted concerns about potential overreliance on RESET and added supervision of students' actions to the robot. To this, two paraprofessionals recounted instances in which students exhibited bullying behaviors toward RESET. This included mocking its speech and issuing rude or dismissive commands "simply to test its limits."

[7] Includes co-design partners (Section 9.3) and all staff who had accompanied or directed a student to the room during any of the three deployment phases.

9.5.6 Case Studies: Example Uses & Shortcomings

P_6, a third-grader diagnosed with ADHD, ASD, Oppositional Defiant Disorder (ODD), and Intellectual Disability (ID), has a history of challenging authority, often requiring extensive staff intervention during de-escalation attempts. His first encounter with RESET was surprising—P_6 engaged independently, responded in full sentences and appropriately followed RESET's prompts with minimal redirection.[8]

In the first week, P_6 frequently requested visits to show RESET his completed classwork. This behavior was not typical for P_6, who previously avoided the de-escalation room and resisted staff-led interventions. Instead of viewing the room as a consequence, P_6 appeared to reframe it as a positive, self-initiated experience. However, his paraprofessional initially questioned whether P_6 was using the robot solely as a social talk companion rather than a self-regulation tool, potentially reinforcing an over-reliance on RESET rather than developing generalizable coping strategies. P_6's paraprofessional later reported leveraging RESET as a neutral third-party mediator, using its small talk to more effectively communicate with P_6 during necessary interventions.

P_{15}, a second-grade student who regularly took teacher-permitted independent breaks, visited RESET to ask clarifying questions about the class lesson on the water cycle. However, due to its observer constraints, RESET maintained its small talk, offering general, non-technical responses rather than the request, direct instruction. As noted by the school psychologist, P_{15} became mildly frustrated by this. She additionally notes how RESET's fixation monitoring detected the sustained focus and adaptively redirected the conversation to a neutral topic about the snacks P_{15} had for lunch. While RESET avoided an overly instructional role by design, it lacked the depth to support such inquiries, relying instead on topic shifts to sustain engagement and prevent fixation.

9.5.7 System Performance

Despite RESET's overall effectiveness, several limitations exist. A key issue was speech recognition errors, where speech-to-text occasionally failed to accurately register students' responses ($N = 10,279$ total), resulting in 8% of instances where users repeated themselves or were misunderstood. Students' reactions varied, ranging from

[8] Video excerpts of both case study interactions and others are included and available at https://youtu.be/ybEuVhxUzhs.

mild frustration to positive engagement with the misunderstanding.

Another challenge was processing latency, where delays in WiFi, speech recognition or response generation disrupted the natural flow of interaction. In 4% of all robot responses (total sample size of $N = 10,814$), noticeable pauses ranging from 2 to 7 seconds caused students to lose focus or abandon interactions prematurely. These responses to latency were rare, and most students patiently tolerated the robot's delay, with some perceiving it as "thinking about what I said" (P_8).

While the system's monitoring appropriately redirected 64% of users' fixated responses ($N = 514$ total; Section 9.4.2), it did not account for when users returned to the fixated topic. Although this strategy aimed to guide more constructive engagement, redirecting can feel dismissive; in 12 instances, students abandoned the conversation entirely, underscoring the need for a more adaptive approach to managing fixation.

9.6 Discussion

In all, the RESET robot integrated well into the school environment, led to more frequent but more efficient use of their de-escalation space, faster cooldown periods, improved transitions back to the classroom, and had lasting effects for how the space was used after the robot's removal.

A common concern with introducing engaging technologies into de-escalation spaces is the risk that students may seek out the space not for regulation, but as a preferred escape from classroom demands. However, RESET's deployment suggests the opposite: despite an increase in overall visit frequency, both visit duration and cooldown times significantly decreased. This means that while students were more likely to visit the de-escalation space during deployment, they also spent less time there, effectively regulating faster and returning to class sooner. If the robot had served as a reinforcing escape mechanism, we would expect students to extend their time in the space rather than reduce it.

At the same time, interaction time with RESET steadily declined over the first three weeks, likely reflecting an initial novelty effect, before stabilizing in the final week. Coupled with shorter overall visits, this trend suggests students gradually relied on RESET less. This challenges the prevailing assumption that robots must maintain high engagement levels to be effective. Instead, this trend suggests that such systems can serve as a short-term regulatory tool that fades in prominence as students internalize self-regulation strategies. This is further reflected in substantially shorter cooldown periods, a trend that persisted even after RESET was removed.

Moreover, RESET helped students move away from passive, individual regulation strategies and instead encouraged the adoption of more socially interactive strategies, such as engaging with others or seeking social support. Importantly, this shift continued even after RESET was no longer actively involved in the de-escalation process, suggesting it had a lasting impact on how students approached self-regulation.

From staff accounts, RESET was viewed as a valuable mediator and peer companion, yet some students challenged it. These instances of robot bullying could reflect a lack of perceived social accountability or attempts to exert control in a structured setting. While such cases were rare in this study, further research is needed to explore how system design can prevent and minimize the reinforcement of bullying behavior.

Our study was conducted in a single ICT school with students, many of whom are neurodiverse based on their IEP status. As such, the findings may not fully generalize to other educational settings or user groups. While our analysis controlled for visit and cooldown times, external factors—such as increased awareness of the de-escalation room, added activities after the robot deployment, and the stress of March's testing period—may have influenced the observed patterns. Future research should explore the robot's performance in diverse school settings, with longer deployment periods and broader student demographics.

9.7 Summary

Amid widespread use of sensory rooms to assist people with emotional and sensory processing challenges, effective implementation remains hindered by limited resources, inconsistent usage, varied user needs, and the risk of reinforcing negative behaviors. In this case, we examined emotional de-escalation—a high-stakes regulatory task—within a public elementary school. Unlike in-home settings where the costs of dysregulation may be more diffuse or private, failure to de-escalate in a school context can lead to missed instructional time, heightened emotional stress, and even physical, social, or disciplinary harm. These contextual factors introduce new challenges and raise critical design considerations for robot-assisted regulation. The system must respond appropriately in real-time, maintain user trust, adapt to fluctuating user needs, engage effectively with both first-time and repeat users, respect the broader social and institutional norms of the setting—all while remaining agnostic to users' age and diagnostic profiles. To address these challenges, we introduce RESET, a socially assistive robot designed to facilitate calming and self-regulatory practices in school de-escalation rooms. Developed iteratively with feedback from key stakeholders and deployed autonomously for one month in a public elementary school, RESET guided students through activities such as deep-breathing exercises, small talk, and collaborative storytelling.

Results from its deployment show that RESET supported smoother transitions back to the classroom, reduced reliance on staff-mediated interventions, and contributed to lasting improvements in students' ability to self-regulate. These were positive effects that persisted even a month after the deployment concluded. Notably, interaction time with RESET declined over the first few weeks before stabilizing, suggesting an initial novelty effect followed by reduced reliance on the robot. This pattern challenges the assumption that sustained engagement[9] is necessary for robotic effectiveness and instead highlights the robot's value as a temporary scaffold that supports the internalization of self-regulation strategies.

This chapter contributes a critical extension of this dissertation's broader goals: designing socially intelligent robots that promote the development of regulation skills through real-world, long-term interactions. In the following chapter, we summarize the key contributions of this dissertation, highlight recurring themes across our studies, and outline avenues of future research.

[9]If inferred from stable or increasing interaction time or similar proxy measures. Relying on interaction time is a common practice in long-term HRI study as discussed in Chapter 2.

Chapter 10

Discussion and Future Directions

Robots, as embodied platforms, offer unique opportunities for on-demand, physically co-present interaction. While the field of robotics has traditionally emphasized reliability and precise motion for physical task assistance, a growing body of literature shows that humans often perceive and engage with robots as social entities. Building on this insight, we explored how robots can provide *social* value and assistance to humans.

To conclude this dissertation, we organize its contributions around a set of central themes. Each theme challenges prevailing assumptions, limitations, or conventions within the field. From these, we propose several directions for future research.

10.1 Central Themes

Our overarching goal is to build intelligent robots that support social regulation therapy. These systems are designed for long-term interaction, operate autonomously in dynamic real-world environments, target novel therapeutic behaviors, and serve users historically underrepresented in the literature.

This dissertation introduces several firsts in the field: the first robots developed specifically for adults with autism; one of the only robotics studies to demonstrate continuous learning progression tied to clinical measures of therapeutic efficacy; the first use of foundation models to deliver unscripted, improvised therapy; and the first robot to address behavioral de-escalation in public spaces while remaining agnostic to users' age or diagnostic profile. We summarize the contributions that extend across multiple chapters of this dissertation.

10.1.1 Long Term Interactions

A central focus of this dissertation is the development of robots that sustain long-term interactions with users. This is because true social learning unfolds over time and requires repeated exposure to novel social situations that test the relevance and adaptability of learned strategies. However, the field remains focused on proof-of-concept studies and feasibility pilots, which tend to prioritize novelty, mere exposure effects, or initial engagement. In order to support the kind of long-term learning required for meaningful gains in social regulation, robots must sustain user engagement over time, move beyond scripted, reactive behaviors toward more proactive and generative interactions, and detect gradual patterns of change in situ. By deploying systems to operate for multiple days or weeks at a time, we create a rich testbed for exploring methods to detect user progress in situ and sustain long-term use.

For instance, Chapters 6 and 8 introduce robots designed to live alongside users in their homes, delivering training experiences that remain relevant, valuable, and engaging over time. As discussed in Chapter 2, sustained user interaction depends on both personalization and adaptation. In line with this, Chapter 8 presents a robot capable of delivering unscripted, personalized, and contextually appropriate interactions tailored to individual users. Chapter 9 further extends this challenge by presenting a deployment context that tests the limits of current personalization methods: a school environment where the robot interacts with a highly diverse user base, ranging from kindergarten to fifth grade students with varied social needs and functioning levels, including both frequent and one-time users.

As we reviewed in Chapter 2, it is common practice for researchers to deploy robots, complete user interactions, and analyze outcomes only after the study has concluded. This retrospective model leaves open important questions about how robots can recognize and adapt to user learning as it unfolds. This consideration is particularly critical in scenarios involving long-term human-robot interaction (HRI). In this dissertation, we revisit the landmark study by Scassellati et al. in 2018 [3], which was the first ever study to explicitly evaluate the generalizability of robot-based autism therapy. The authors assessed clinical skill transfer: whether the behaviors children learned during robot interactions transferred to a human partner in the absence of the robot. Although this approach spoke to the clinical efficacy of the month-long intervention, these were isolated assessments that captured user behavior before the therapy and after the therapy. In Chapter 5, we extend this line of inquiry by exploring how robots can continuously track learning progression in situ. Given the vast heterogeneity of autism, the unpredictability of children behavior, and the challenges of accurate robot perception in home environments, we required robust computational methods to automatically detect and interpret behavioral change over time. Chapter 5 illustrates how robots deployed for long-term HRI can capture fine-grained, continuous indicators of individual learning progression, mapped to clinically validated measures of therapeutic efficacy.

Furthermore, while much attention in robot design is devoted to initial user engagement, the offboarding process (how a robot exits the user's life after the intervention ends) is equally important. When we approach building robots for long-term interactions, we must recognize that relationships users form with robots can carry significant emotional weight. In our work, we treat the entire deployment pipeline—including introduction to the robot, its physical setup, in-situ troubleshooting or maintenance, exit strategies, and offboarding—as a series of essential design considerations (e.g., Chapters 6, 8, 9). These design considerations become even more critical when robots are intended to live alongside users in their real-world, everyday personal environments. We elaborate on this below.

10.1.2 In Dynamic, Real-World Environments

All of the work presented in this dissertation takes place outside of the controlled laboratory environment, occurring instead in users' everyday personal spaces, where interactions are minimally constrained and designed to be highly adaptable and personalized. While these settings offer greater ecological validity and relevance, they also introduce a wide range of technical challenges that are not typically encountered in lab-based studies. Real-world environments are dynamic and unstructured: lighting conditions vary throughout the day, background noise fluctuates, physical layouts vary dramatically across deployment sites, and users engage with the robot amidst competing demands, interruptions, and distractions.

These real-world contexts introduce significant social, physical, and organizational variability, requiring the robot to respond flexibly to unstructured and unpredictable conditions. From a technical standpoint, this creates challenges across nearly every system layer. Perception systems must operate reliably despite environmental noise, poor lighting, or camera occlusion. Speech recognition must adapt to different speaking styles and ambient conditions. On the decision-making side, the robot must continuously assess user state, interaction history, and contextual cues to determine when to act, when to wait, and how to adjust its strategy. Additionally, safety, privacy, and autonomy must be maintained without relying on real-time human oversight—meaning the robot must not only act appropriately, but must also know when *not* to act.

This brings forward new questions around social appropriateness. When is it the right moment to engage a user in an interaction? When should the robot remain silent, offer encouragement, or redirect attention? Unlike scripted lab tasks, real-world interactions demand that the robot make context-sensitive decisions about *if*, *when*, and *how* to intervene. As we show in Chapters 8 and 9, addressing these challenges requires not only robust sensing and autonomy, but also mechanisms for real-time behavioral judgment—ensuring the robot's actions are not only functional, but socially attuned and appropriate for the user's context and state.

10.1.3 Fully Autonomous Robot Operation

All of the robots across our studies were designed for fully autonomous operation. Achieving this required overcoming substantial technical challenges in system architecture, perception, and decision-making. These robots had to interpret noisy sensor input, detect changes in user state or context, and respond in socially appropriate ways—all in real time, for extended deployments, and without human oversight.

Many social regulation skills are learned implicitly and vary contextually. Because these behaviors are not governed by fixed rules and are rarely taught through explicit instruction, they are not easily scripted or pre-programmed. Systems that rely on rigid rule-based approaches can produce interactions that are brittle, unnatural, or short-lived. To address this, our robots must first be capable of simulating or modeling the target behavior, either to convey its appropriate expression or to effectively prompt it in users (Chapters 5, 6, 8, 9). They must also recognize when user behaviors align with desired outcomes in real time (Chapters 5, 8, 9), and crucially, infer when and how to respond, reinforce, or give feedback to support continued learning and engagement (Chapters 7–9).

Moreover, social regulation depends on internal emotional and cognitive states (e.g., frustration, anxiety, attention) that are not directly observable. Inference must occur through noisy proxies like gaze, latency, speech patterns, or physiological data—each with limited reliability and especially fragile under real-world or individual user variation. While extensive research has focused on developing reliable off-the-shelf models for automated user behavior detection, we frequently encountered limitations when applying these models to our specific user populations and deployment contexts.

For example, gaze estimation models trained on neurotypical adults often failed to generalize to children with autism, whose gaze behavior may be atypical (Chapter 5). In-home detection systems struggled with false positives due to the presence of human-like faces on televisions, toys, or images (Chapters 6 and 8). Similarly, speech transcription became unreliable when the robot must distinguish between user-directed speech and ambient dialogue from other people or media sources (motivating our development of detection models applied in Chapters 6, 8, and 9, e.g., [34]). In the absence of reliable off-the-shelf perception models, our systems involve hybrid approaches that combine lightweight heuristics, contextual rules, and adaptive thresholds tailored to the deployment environment (e.g., our grounded observer framework presented in Chapter 7 and applied in Chapters 8 and 9).

10.1.4 Novel Behavioral Targets for Therapy

While prior work in socially assistive robotics has primarily focused on foundational intervention targets—such as joint attention, imitation, and labeling or recognizing emotions—this dissertation explores more nuanced and socially embedded behavioral goals that have received less attention in the field. These novel targets reflect challenges users face in real-world, dynamic environments and require more sophisticated forms of social reasoning and adaptability from robotic systems. For instance, in Chapter 6, we examine how robots can support resilience to everyday interruptions—a critical, yet often overlooked, aspect of social regulation. Rather than training users on discrete social behaviors in isolation, the system helps individuals practice recovering and refocusing after disruptions, promoting the learning of regulation strategies in the context of daily life. In Chapter 8, we explore small talk as an intervention target. While seemingly simple, small talk encompasses a range of core social skills, including conversational turn-taking, topic maintenance, perspective-taking, emotion sharing, and social timing. Both small talk and interruptions resiliency are examples of skills that are not directly addressed in traditional therapeutic contexts, yet our studies find they are tied to long-term life outcomes and are essential for social inclusion and daily functioning.

10.1.5 Understudied Users in Unique Contexts

This body of work expands the reach of assistive robots to populations and contexts that are understudied in HRI research. For example, despite decades of progress in autism research, the vast majority of studies and clinical programs continue to focus almost exclusively on children. Therefore, little is known about the specific needs of adults or how best to support positive outcomes in adult life. In Chapter 8, we begin by openly exploring what skills are needed and valued by adults with autism based on their lived experiences. We then organize their input into a structured framework (small talk training) for practicing social skills in a way that reflects their identified goals and priorities. Our work in Chapters 6 and 8 presents the first ever in-home robots designed specifically for adults with autism.

We also investigate how robots can support children facing emotional and behavioral regulation challenges in complex public school environments. In Chapter 9, we introduce RESET, a socially assistive robot deployed in a school de-escalation room. RESET interacted with students across a wide age range and with diverse behavioral and developmental profiles. Some students engaged with the robot frequently, while others encountered it only once, requiring the system to be flexible and effective across a broad spectrum of user needs and interaction frequencies. This context challenged traditional models of personalization, demonstrating how robots can deliver contextually appropriate support without reliance on pre-specified diagnostic categories or rigid behavioral assumptions.

The COVID-19 pandemic marked an unprecedented period in modern history—billions of individuals worldwide were confined to their homes under emergency health and social distancing mandates. This prolonged isolation highlighted several areas of social-emotional health that can benefit from therapuetic support. In response to these challenges, we developed a robot teleoperation system that allowed users to control and communicate through a robot located in a peer's home, enabling children to engage in physical play and social interaction despite geographic separation (Chapter 4). More broadly, the majority of our studies were conducted under COVID-19 safety protocols (Chapters 4, 6, 8, 9). These constraints directly informed our system design: robots were built for contactless delivery, designed to be easily set up by participants without technical expertise, intuitive to use, and capable of operating independently without ongoing maintenance or researcher oversight. This emphasis on autonomy and accessibility was essential not only for maintaining safety but also for ensuring feasibility and scalability in real-world deployments during and beyond the pandemic.

Humans differ widely in their developmental trajectories, interaction styles, personalities, preferences, and cognitive functioning—especially within highly heterogeneous populations such as individuals with autism. This variability presents both a design and modeling challenge: robots must operate flexibly without relying on uniform behavioral baselines or one-size-fits-all interaction patterns. Our approach to this is reflected in iterative design methodologies in which we collaborate directly with specialized populations to understand their needs and inform design objectives (e.g., in Chapters 6, 8, and 9). In practice, we developed systems that operate without requiring individualized pre-training, instead adapting though behavior trees or symbolic overlays that adjust to observed user behavior in real-time (e.g., Chapter 7), robust default strategies to function reasonably across a wide range of behaviors (e.g., [33, 34] applied in Chapters 6–9), and guardrails that constrain generative outputs to ensure safety and appropriateness in novel, unanticipated scenarios (Chapter 7–9).

10.2 Directions for Future Research

This dissertation examines how we can design, develop, and deploy robots to support sustained social regulation. Our studies present ways in which robots can be tailored to specialized user needs, embed opportunities for therapy in naturalistic real-world spaces, and can function independently and reliably for long-term interactions. In the continued pursuit of this goal, we discuss opportunities for further research.

10.2.1 When Robots Should Break the Rules

Humans hold several expectations about robots and these expectations are reflected in the ways that researchers build robots. This section outlines seven common expectations frequently embedded in robot design and presents a case for why, in certain contexts, violating these expectations can lead to more effective, ethical, or socially intelligent behavior. For each expectation, we briefly describe scenarios in which deliberately breaking the rule can lead to better social outcomes.

Rule 1: Robots Should Always Be Willing to Engage

Robotics research often features systems that are always on—constantly aware, attentive, and ready to interact. This availability may be expressed through proactive behaviors, such as autonomously tidying clutter in a home or navigating a facility to collect environmental data, or through reactive responses, like answering user questions or responding to a wave or voice prompt. These examples reflect a common design assumption: that social robots should remain perpetually "awake," always ready to engage with either users or the surrounding environment.

However, continuous availability is not always optimal—and in many real-world contexts, it can be socially inappropriate, cognitively exhausting, or simply unwelcome. There are moments when it is more respectful or effective for a robot to power down, go to "sleep," or deliberately ignore interaction attempts. For instance, in emotionally charged moments, a robot that remains silent rather than intervening may help a user regain composure or preserve a sense of privacy. In classroom settings, ignoring a student's off-topic whisper can prevent unnecessary distraction. Similarly, when users test boundaries by issuing inappropriate or repetitive prompts, selective non-responsiveness can serve as a form of behavioral shaping, discouraging misuse while reinforcing more appropriate engagement patterns. Robots may also need to ignore low-priority social bids in order to focus on more urgent tasks, or respect socio-cultural contexts where silence is expected. In light of this, future work should explore how robots might learn when to strategically ignore interactions with users.

This becomes especially salient in long-term deployments where robots share living spaces with users for extended periods. For instance, in our own deployments (e.g., Chapters 6 and 8), robots remain in users' homes for several days. Just as it would be inappropriate for a human therapist to enter someone's home uninvited and announce that it is time for therapy, it is likely problematic for a robot to do so. Accordingly, researchers should equip robots with the ability to determine when they should refrain from initiating or responding to interactions.

Rule 2: Robots Should Always Offer Help

Helping is a fundamental dynamic of human-human relationships. Yet, the act of offering help can sometimes be met with resistance [607, 608]. For example, what goes on and what goes wrong when one volunteers to help a friend and is rudely rebuffed? It has been said that the word "help" itself comes up primarily when someone is described to have "not been helpful" [607]. Robots are built to assist people. In contrast to human-human dynamics, it is generally assumed that robots should be readily available when needed and always willing to help its users [609–611]. We challenge this prevailing paradigm that robots should inherently always offer assistance.

There are many situations in which a robot should opt to withhold offering help, even when it is technically capable of assisting. For instance, in therapeutic or rehabilitation contexts, withholding help can encourage independence—such as when a robot observes a user struggling slightly to stand but allows them to complete the motion on their own to support recovery goals. A robot in a restaurant observes a server dropping a fork but refrains from offering help, recognizing that stepping in would interrupt the flow of professional service and draw attention. In group settings, a robot may refrain from jumping in with an answer to preserve conversational flow or give someone a chance to recall information independently. Robots may also withhold assistance when user preferences are known (e.g., a user who prefers manual control over cooking tasks) or when the context is ambiguous and premature intervention could cause confusion or offense. In these cases, not helping is not a limitation of the robot, but a strategic behavior aligned with social, emotional, or pedagogical goals. Future work should examine how robots can discern when it is appropriate to offer or withhold assistance to users.

Rule 3: Robots Should Always Be Task-Productive

Robots are often evaluated by their efficiency and task-oriented success. These measures are not limited to functional task performance. For instance, some studies define success as increasing the amount of eye contact users make with the robot, treating it as a proxy for engagement. Conversational therapy systems may aim to maximize speaking time as a stand-in for user comfort, while service robots often optimize for metrics like task completion time or the number of customers served. These benchmarks reflect broader societal values that emphasize output, speed, and optimization. However, in many social and collaborative contexts, rigid adherence to task goals may inadvertently undermine relational dynamics or overlook the importance of small, seemingly "unproductive" moments that contribute to trust, rapport, and long-term acceptance. For example, should factory assembly lines include robots capable of small talk with its human collaborators? While it may not always translate directly to short-term task outcomes, such interactions could foster a more positive work environment, reduce stress, and support human well-being. Future work should reconsider what it means for a robot to be successful, expanding evaluation criteria to include social value and relational outcomes, not just efficiency.

Rule 4: Robots Should Always Be Polite and Deferential

We typically design robots to be polite [612, 613]. Researchers incorporate system-level rules to avoid potentially interrupting, contradicting, or confronting users, as a way of allowing the robot to signal friendliness and minimize social friction. However, politeness and deference can be counterproductive. For example, consider a rule that restricts the robot from interrupting a user while they are speaking—a generally sound and polite constraint. However, in a situation where the user begins to spiral into a repetitive or self-deprecating monologue, the rule may need to be relaxed to allow a well-timed, gentle interruption that redirects or re-engages the user constructively. In this case, the rule's intent (respecting user agency) must be weighed against its current utility and possible harm.

In educational settings, a robot tutor may need to interrupt a student mid-explanation to correct a fundamental misunderstanding before it becomes entrenched. In healthcare, a robot reminding a patient about medication adherence may need to persist or escalate its tone if polite prompting is repeatedly ignored. Even in customer service, a robot may need to push back gently when a user makes an unreasonable request, such as asking it to perform actions outside its scope or to behave inappropriately. In these cases, assertiveness is not a breakdown in politeness but rather a demonstration of situational awareness and a commitment to supporting human goals responsibly.

Future research should explore how robots can balance politeness with assertiveness, developing context-aware strategies that allow them to interrupt, redirect, or disagree when doing so supports user safety, learning, or long-term well-being.

Rule 5: Robots Should Never Withhold Information or Lie

It is natural to expect that robots should always provide complete and accurate information when asked. This expectation reflects a view of robots as transparent, factual tools designed to reduce uncertainty and deliver immediate answers. However, in socially and ethically complex situations, unconditional disclosure can be inappropriate or even dangerous. For example, in elder care contexts, a robot supporting a person with dementia may know where the car keys are but choose to withhold that information if there is reason to believe the person may attempt to drive unsafely or leave the house without supervision. In this case, withholding is a protective measure that prioritizes the user's physical safety over immediate compliance. Similarly, a therapeutic robot may avoid answering certain personal questions if doing so could trigger distress, or delay factual responses to encourage problem-solving in educational settings.

There are also situations where providing partial information may be more appropriate than full disclosure. For example, a healthcare robot may inform a patient that their test results are being reviewed, without immediately sharing abnormal findings, allowing a physician to deliver the results in a controlled clinical context. In extreme cases, strategic deception may be ethically justified—for example, a robot may state that the building exits are temporarily inaccessible during a lockdown to prevent individuals from moving toward danger. These examples demonstrate that rigid truth-telling can be socially and ethically insufficient. Future work should examine how robots can make context-sensitive decisions about when to disclose, withhold, or modulate information in ways that prioritize safety, well-being, and appropriate delegation of sensitive communication.

Rule 6: Robots Should Never Make Mistakes

Robots are often designed to appear competent, consistent, and error-free—reflecting the belief that reliability and precision are core to their value as machines. As a result, mistakes are typically treated as design flaws to be avoided or corrected. However, in social contexts, occasional and intentional errors can serve important relational and pedagogical functions. For example, in educational settings, a robot that makes a simple mistake while solving a problem may prompt the user to step in and correct it—an interaction that reinforces learning through a "learning-by-teaching" paradigm. In other cases, a robot might lose a game on purpose to boost a child's confidence or encourage continued engagement. Small, human-like errors can also build rapport by making the robot seem more relatable, fallible, and less intimidating. These behaviors, when used deliberately and transparently, can signal humility, invite user participation, and foster trust. Future research should examine how robots can strategically use mistakes to support social, emotional, and learning outcomes without undermining overall user confidence in the system.

Rule 7: Robots Should Never Model Harmful Behavior

Robots are typically designed to avoid behaviors that may be interpreted as aggressive, exclusionary, or morally inappropriate—such as mocking, taunting, or bullying. These behaviors are widely considered unacceptable in human interaction, and by extension, are excluded from robot conduct to maintain trust and psychological safety. However, in controlled settings, robots can strategically model norm-violating behavior to promote reflection, learning, and prosocial action. For instance, prior work has used two robots to simulate a bullying scenario, where one robot teases or excludes the other, to study how children respond as bystanders [399]. These scenarios are designed not to normalize bullying, but to prompt users to recognize mistreatment and practice appropriate intervention strategies. By witnessing norm violations enacted by robots, users are given a safe, repeatable context in which to explore empathy, fairness, and the moral imperative to speak up.

10.2.2 Rethinking the Intelligence Robots Need to Deliver Therapy

To achieve our goal of building intelligent robots for social regulation therapy, we first had to examine how to build robots capable of regulating their own social behavior. This idea presents a promising avenue for future research, which we discuss below.

Humans communicate in ways that are inherently contextual, socially situated, and often ambiguous. For robots to participate meaningfully in such interactions, they must be able to interpret and respond to this complexity. Historically, the prevailing approach in HRI research has relied on simple rule-based systems that map specific human behaviors to predefined robot actions, leading to heavily scripted and inflexible interactions. Although these systems allow for more controlled experimentation, they tend to be brittle and difficult to generalize across diverse social contexts. Yet, a majority of HRI studies continue to rely on rule-based, scripted frameworks. In contrast, our work leverages the high-dimensional representational space of foundation models to enable more flexible and adaptive robot behavior (as demonstrated in Chapters 7–9).

These models are large-scale statistical models trained on massive datasets, and their internal decision-making processes are often opaque. Consequently, their outputs are not strictly deterministic (especially in cases involving online adaptation or human-in-the-loop fine-tuning) and can range from highly accurate and contextually appropriate to factually incorrect, irrelevant, or synthetically generated content not grounded in any source data (i.e., "hallucinations"). Given these limitations, it would be ethically inappropriate to deploy foundation models on physically embodied robot platforms for direct user interaction. More critically, within the broader aims of this dissertation, it would be unreasonable to do so in the context of unsupervised, autonomous operation with vulnerable users in their personal environments over extended periods of time.

Recently, there have been substantial efforts to create guardrails for foundation models [501, 614]. However, most of these effort focus on disembodied systems for task-oriented assistance, where success is typically measured by clear, quantifiable performance metrics. The work presented in this dissertation is among the first to explore how behavioral constraints can be enforced in physically embodied systems that interact socially with humans. We developed a mechanism for establishing robust guardrails on foundation models (i.e., the grounded observer, Chapter 7). We then demonstrated our mechanism to be effective for enabling flexible small talk (Chapters 7–9), delivering personalized feedback on social skills practice (Chapter 8), and seamlessly transitioning users to new therapeutic activities (Chapter 9). This resulted in the first socially assistive robots capable of delivering therapy through unscripted, spontaneous, and improvised interactions with users. Building on these advances, we can now envision novel methods of therapy.

Adjusting Therapy Structure in Real Time

As robots gain the ability to deliver more flexible and context-sensitive interactions, new opportunities emerge for them to actively guide and adapt therapeutic experiences in real time. In a home environment, a robot helping a child practice conversation initiation might begin in a simple one-on-one format. Once the child demonstrates sufficient fluency, the robot can strategically shift to a triadic configuration by involving other family members—such as prompting a sibling to join a game or encouraging the child to ask a parent a question—thereby embedding the therapy more deeply in the user's immediate social and real-world context. The ability to flexibly adjust the overall intervention structure is not only valuable for scaling interventions but also serves as a natural test of generalization. It allows the robot to observe whether the learned behavior persists when directed toward other humans outside of the specific human-robot dyad.

In addition to moving from dyadic to triadic or group configurations, robots can adapt the interaction structure in numerous other meaningful ways. For instance, in classroom activities where multiple students work together, an observer-enabled robot (see Chapter 7) may initially enforce a rule that prioritizes equal turn-taking among children. However, if one child becomes visibly dysregulated (crying, shutting down, or isolating), the robot may need to override this fairness constraint to focus its attention on that individual. Temporarily suspending the group-level turn-taking rule allows the system to attend to a more urgent emotional need. After dyadic intervention that support the child's recovery, the robot can gradually shift its behavior to support the child's reintegration into the group. This might involve inviting peers to reengage the child through a cooperative task, assigning the child a small leadership role to restore a sense of agency, or subtly reshaping the group dynamic to create a more inclusive atmosphere.

By leveraging the contextual awareness and generative flexibility of foundation models, robots can make real-time decisions about when to introduce new interaction structures or participants to both support continued learning and evaluate skill generalization in more ecologically valid ways.

Dynamically Revising Goals, Strategies, and Content

Therapeutic interventions often focus on a fixed set of behavioral targets (e.g., making eye contact, initiating conversation). However, as users progress, these initial targets may become outdated, overly simplistic, or misaligned with emerging needs. We can envision robots that can dynamically revise these targets based on observed patterns of mastery, emerging challenges, or contextual demands.

For instance, a robot supporting an adult undergoing social anxiety therapy might initially focus on brief verbal initiations (e.g., saying hello or answering yes/no questions). As the user gains confidence and fluency, the robot could gradually scaffold more complex behaviors, such as asking more open-ended questions. It might also increase the variability or unpredictability of its responses to better simulate real-world interactions and reduce over-reliance on rehearsed scripts. Moreover, if new challenges arise—such as the user fixating on a recent emotional experience or struggling to shift attention—the robot can reprioritize its therapeutic goals and introduce targeted support for narrative coherence and emotion regulation (as we demonstrated in Chapter 9).

For more ambitious long-term systems—those designed to interact autonomously with users over the course of months or even years—this capacity to dynamically revise therapeutic goals and strategies means users can experience more developmentally appropriate support. Robots can generate novel curricula and content that remain engaging, personalized, and contextually relevant over time. We can envision systems that grow alongside users and offer sustained lifelong support.

10.2.3 Novel Users, Spaces, and Skills

The central themes outlined in Sections 10.1.1–10.1.5 each represent distinct and important directions for future research. It is a necessary branch of future research to explore how robots that can sustain user engagement spanning even longer timelines. Future work can explore additional real-world settings and the distinct logistical, ethical, and social demands each context places on robotic systems. Environments such as foster or transitional housing, youth detention centers, and homeless shelters remain absent from the current literature. Yet, these spaces represent critical contexts where socially assistive robots could deliver meaningful, high-impact support. A wide range of social behaviors continue to pose challenges for autonomous robotic systems; we proposed examples of such behaviors and the nuanced judgment they require (Section 10.2.1). There are many underserved populations and underexplored contexts where socially assistive systems could offer meaningful support. For example, how can we design robots that help mitigate anxiety for patients actively undergoing chemotherapy? How can we build systems that support veterans coping with post-traumatic stress? How might robots assist teens struggling with eating disorders? What unique design challenges must be addressed when creating systems intended to support users experiencing clinical depression?

10.2.4 Ethical Considerations

In this dissertation, we propose building intelligent robots for social regulation therapy. In doing so, we see the necessary requirement for robots to interact autonomously with vulnerable users in their personal, everyday settings for extended periods of time. The design and deployment of these systems therefore raise critical ethical questions.

First, as researchers, we are drawn to robots as tools for supporting human outcomes precisely because of their potential to influence human social behavior. However, this very capacity raises a critical ethical question: *how much influence is too much?* At what point does helpful guidance become coercion, or support become manipulation? As we continue to build robots that socially interact with humans, we must carefully consider the ethical boundaries of their influence.

Key ethical challenges in the design and deployment of assistive robots include safeguarding user autonomy, ensuring transparency in robot decision-making, and protecting the privacy and dignity of users. As robots become more adaptive and emotionally responsive, there is a growing risk that users may misunderstand the robot's capabilities, assign it undue authority, or develop inappropriate levels of trust or attachment. Designers must be vigilant in preventing scenarios where users feel manipulated, surveilled, or emotionally misled.

To this point, maintaining user autonomy requires careful attention to how robots initiate interaction, guide behavior, and influence decision-making—especially when working with populations that may have reduced cognitive or communicative capacity. Transparency must extend beyond the system's actions to include its limitations; users and caregivers should have a clear understanding of what the robot can and cannot do, how decisions are made, and when human intervention is necessary.

Informed consent poses unique challenges in this domain. For users with limited ability to understand the purpose or function of the system (such as young children, individuals with developmental disabilities, or those in states of emotional distress), obtaining ongoing, meaningful consent may require multimodal communication, caregiver involvement, and adaptive consent strategies that evolve alongside the user (i.e., consent should not be treated as a one-time event). Moreover, robots operating in shared spaces must consider the rights of bystanders who may influenced by the system or whose data or interactions may be incidentally captured.

Finally, future work must attend to the long-term social effects of widespread deployment and adoption. While socially assistive robots may offer immediate benefits (such as improved access to therapy or a sense of companionship), their presence also has the potential to subtly reshape how we relate to others, what we expect from care, and how we express emotional needs. For example, if a robot consistently offers nonjudgmental attention or immediate feedback, will users begin to prefer robotic interaction over more complex, less predictable human relationships? Will frequent exposure to emotionally calibrated responses from robots shift how people communicate distress or seek support from others? In summary, ethical reflection must be embedded not only at the point of deployment, but throughout the lifecycle of system design, from early prototyping to long-term field use.

Bibliography

[1] M. Paetzel, G. Perugia, and G. Castellano, “The persistence of first impressions: The effect of repeated interactions on the perception of a social robot,” in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 73–82.

[2] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro, “A two-month field trial in an elementary school for long-term human–robot interaction,” *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 962–971, 2007.

[3] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic, “Improving social skills in children with ASD using a long-term, in-home social robot,” *Science Robotics*, vol. 3, no. 21, p. eaat7544, 2018.

[4] J. Fasola and M. J. Matarić, “A socially assistive robot exercise coach for the elderly,” *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 3–32, 2013.

[5] D. Silvera-Tawil, D. Bradford, and C. Roberts-Yates, “Talk to me: The role of human-robot interaction in improving verbal communication skills in students with autism or intellectual disability,” in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 1–6.

[6] A. K. Ostrowski, C. Breazeal, and H. W. Park, “Mixed-method long-term robot usage: Older adults’ lived experience of social robots,” in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 33–42.

[7] E. Hyun, H. Yoon, and S. Son, “Relationships between user experiences and children’s perceptions of the education robot,” in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2010, pp. 199–200.

[8] S. Jeong, L. Aymerich-Franch, S. Alghowinem, R. W. Picard, C. L. Breazeal, and H. W. Park, “A robotic companion for psychological well-being: A long-term investigation of companionship and therapeutic alliance,” in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2023, pp. 485–494.

[9] C. E. Weinstein, J. Husman, and D. R. Dierking, "Self-regulation interventions with a focus on learning strategies," in *Handbook of Self-Regulation*. Elsevier, 2000, pp. 727–747.

[10] B. J. Zimmerman, "Attaining self-regulation: A social cognitive perspective," in *Handbook of Self-Regulation*. Elsevier, 2000, pp. 13–39.

[11] A. Bandura and R. H. Walters, *Social learning theory*. Prentice Hall, Englewood Cliffs, New Jersey, USA, 1977, vol. 1.

[12] J. J. Gross, "Emotion regulation: Current status and future prospects," *Psychological Inquiry*, vol. 26, no. 1, pp. 1–26, 2015.

[13] N. Eisenberg, T. L. Spinrad, and N. D. Eggum, "Emotion-related self-regulation and its relation to children's maladjustment," *Annual Review of Clinical Psychology*, vol. 6, no. 1, pp. 495–525, 2010.

[14] C. McKown, L. M. Gumbiner, N. M. Russo, and M. Lipton, "Social-emotional learning skill, self-regulation, and social competence in typically developing and clinic-referred children," *Journal of Clinical Child & Adolescent Psychology*, vol. 38, no. 6, pp. 858–871, 2009.

[15] C. A. Mazefsky and S. W. White, "Emotion regulation: Concepts & practice in autism spectrum disorder," *Child and Adolescent Psychiatric Clinics of North America*, vol. 23, no. 1, pp. 10–1016, 2013.

[16] L. S. Caddell and L. Clare, "The impact of dementia on self and identity: A systematic review," *Clinical Psychology Review*, vol. 30, no. 1, pp. 113–126, 2010.

[17] C. Knight and L. D. Borders, "Trauma-informed supervision: Core components and unique dynamics in varied practice contexts," *The Clinical Supervisor*, vol. 37, no. 1, pp. 1–6, 2018.

[18] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer, "Understanding social robots," in *2009 Second International Conferences on Advances in Computer-Human Interactions*. IEEE, 2009, pp. 169–174.

[19] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social robotics," *Springer Handbook of Robotics*, pp. 1935–1972, 2016.

[20] B. Scassellati, H. Admoni, and M. Matarić, "Robots for use in autism research," *Annual Review of Biomedical Engineering*, vol. 14, no. 1, pp. 275–294, 2012.

[21] P. Pennisi, A. Tonacci, G. Tartarisco, L. Billeci, L. Ruta, S. Gangemi, and G. Pioggia, "Autism and social robotics: A systematic review," *Autism Research*, vol. 9, no. 2, pp. 165–183, 2016.

[22] K. Matheus, R. Ramnauth, B. Scassellati, and N. Salomons, “Long-term interactions with social robots: Trends, insights, and recommendations,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 14, no. 3, pp. 1–42, 2025.

[23] R. A. Thompson, “Emotion regulation: A theme in search of definition,” *Monographs of the Society for Research in Child Development*, pp. 25–52, 1994.

[24] P. M. Cole, S. E. Martin, and T. A. Dennis, “Emotion regulation as a scientific construct: Methodological challenges and directions for child development research,” *Child Development*, vol. 75, no. 2, pp. 317–333, 2004.

[25] J. J. Gross, “The emerging field of emotion regulation: An integrative review,” *Review of General Psychology*, vol. 2, no. 3, pp. 271–299, 1998.

[26] J. J. Gross and O. P. John, “Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being.” *Journal of Personality and Social Psychology*, vol. 85, no. 2, p. 348, 2003.

[27] K. N. Ochsner and J. J. Gross, “The cognitive control of emotion,” *Trends in Cognitive Sciences*, vol. 9, no. 5, pp. 242–249, 2005.

[28] T. L. Webb, E. Miles, and P. Sheeran, “Dealing with feeling: a meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation.” *Psychological Bulletin*, vol. 138, no. 4, p. 775, 2012.

[29] R. Ramnauth, E. Adéníran, T. Adamson, M. A. Lewkowicz, R. Giridharan, C. Reiner, and B. Scassellati, “A social robot for improving interruptions tolerance and employability in adults with ASD,” in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 4–13.

[30] R. Ramnauth, D. Bršić, and B. Scassellati, “A robot-assisted approach to small talk training for adults with ASD,” in *Proceedings of the Robotics: Science and Systems Conference (RSS)*, Jun. 2025.

[31] R. Ramnauth, D. Brščić, and B. Scassellati, “From fidgeting to focused: Developing robot-enhanced social-emotional therapy (RESET) for school de-escalation rooms,” in *Proceedings of the 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2025.

[32] ——, “A grounded observer framework for establishing guardrails for foundation models in socially sensitive domains,” *arXiv preprint arXiv:2412.18639*, 2024.

[33] ——, “Should I help?: A skill-based framework for deciding socially appropriate assistance in human-robot interactions,” in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2024, pp. 2051–2058.

[34] N. C. Georgiou, R. Ramnauth, E. Adeniran, M. Lee, L. Selin, and B. Scassellati, "Is someone there or is that the TV? detecting social presence using sound," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 12, no. 4, pp. 1–33, 2023.

[35] R. Ramnauth, F. Shic, and B. Scassellati, "Gaze behavior during a long-term, in-home, social robot intervention for children with ASD," in *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, ser. HRI '25. IEEE Press, 2025, p. 949–957.

[36] N. Tsoi, J. Connolly, E. Adéníran, A. Hansen, K. T. Pineda, T. Adamson, S. Thompson, R. Ramnauth, M. Vázquez, and B. Scassellati, "Challenges deploying robots during a pandemic: An effort to fight social isolation among children," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021, pp. 234–242.

[37] J. Saldien, K. Goris, B. Vanderborght, J. Vanderfaeillie, and D. Lefeber, "Expressing emotions with the social robot Probo," *International Journal of Social Robotics (IJSR)*, vol. 2, pp. 377–389, 2010.

[38] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 82–90.

[39] S. Sabanovic, M. P. Michalowski, and R. Simmons, "Robots in the wild: Observing human-robot social interaction outside the lab," in *9th IEEE International Workshop on Advanced Motion Control, 2006.* IEEE, 2006, pp. 596–601.

[40] I. Leite, A. Pereira, S. Mascarenhas, C. Martinho, R. Prada, and A. Paiva, "The influence of empathy in human–robot relations," *International Journal of Human-Computer Studies*, vol. 71, no. 3, pp. 250–260, 2013.

[41] M. Bajones, D. Fischinger, A. Weiss, P. D. L. Puente, D. Wolf, M. Vincze, T. Körtner, M. Weninger, K. Papoutsakis, D. Michel *et al.*, "Results of field trials with a mobile service robot for older adults in 16 private households," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 2, pp. 1–27, 2019.

[42] A. Weiss, A. Pillinger, and C. Tsiourti, "Merely a conventional "diffusion" problem? on the adoption process of Anki Vector," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 712–719.

[43] A. Barco, J. Albo-Canals, C. Garriga-Berga, X. Vilasís-Cardona, L. Callejón, M. Turón, C. Gómez, and A. López-Sala, "A drop-out rate in a long-term cognitive rehabilitation program through robotics aimed at children with TBI,"

in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2014, pp. 186–192.

[44] Z. Zhao and R. McEwen, "“let's read a book together”: A long-term study on the usage of pre-school children with their home companion robot," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 24–32.

[45] A. M. Rosenthal-von der Pütten, A. Weiss, and S. Šabanović, "The challenge (not) to go wild! Challenges and best practices to study HRI in natural interaction settings," in *The 11th ACM/IEEE International Conference on Human Robot Interaction (HRI)*. IEEE Press, 2016, p. 583–584.

[46] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: a survey," *International Journal of Social Robotics (IJSR)*, vol. 5, pp. 291–308, 2013.

[47] J. M. Kory-Westlund, H. Won Park, I. Grover, and C. Breazeal, "Long-term interaction with relational SIAs," in *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, 2022, pp. 195–260.

[48] A. Ramachandran, S. S. Sebo, and B. Scassellati, "Personalized robot tutoring using the assistive tutor POMDP (AT-POMDP)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8050–8057.

[49] M. Donnermann, P. Schaper, and B. Lugrin, "Investigating adaptive robot tutoring in a long-term interaction in higher education," in *2022 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 171–178.

[50] A. Jones and G. Castellano, "Adaptive robotic tutors that support self-regulated learning: A longer-term investigation with primary school children," *International Journal of Social Robotics (IJSR)*, vol. 10, pp. 357–370, 2018.

[51] A. Afyouni, I. Ocnarescu, I. Cossin, E. Kamoun, A. Mazel, and C. Fattal, "Living one week with an autonomous Pepper in a rehabilitation center: lessons from the field," in *2022 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 554–559.

[52] C. D. Kidd and C. Breazeal, "Robots at home: Understanding long-term human-robot interaction," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 3230–3235.

[53] I. P. Bodala, N. Churamani, and H. Gunes, "Teleoperated robot coaching for mindfulness training: A longitudinal study," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 939–944.

[54] I. Leite, M. McCoy, M. Lohani, D. Ullman, N. Salomons, C. Stokes, S. Rivers, and B. Scassellati, "Emotional storytelling in the classroom: Individual versus group interaction between children and robots," in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 75–82.

[55] J. K. Westlund, G. Gordon, S. Spaulding, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Learning a second language with a socially assistive robot," *Almere, The Netherlands*, 2015.

[56] D. P. Davison, F. M. Wijnen, V. Charisi, J. van der Meij, V. Evers, and D. Reidsma, "Working with a social robot in school: a long-term real-world unsupervised deployment," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 63–72.

[57] T. Q. Trinh, A. Vorndran, B. Schuetz, B. Jaeschke, A. Mayfarth, A. Scheidig, and H.-M. Gross, "Autonomous mobile gait training robot for orthopedic rehabilitation in a clinical environment," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 580–587.

[58] B. Irfan, N. C. Gomez, J. Casas, E. Senft, L. F. Gutiérrez, M. Rincon-Roncancio, M. Munera, T. Belpaeme, and C. A. Cifuentes, "Using a personalised socially assistive robot for cardiac rehabilitation: A long-term case study," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 124–130.

[59] R. Ramnauth, E. Adéníran, T. Adamson, M. A. Lewkowicz, R. Giridharan, C. Reiner, and B. Scassellati, "A social robot for improving interruptions tolerance and employability in adults with ASD," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 4–13.

[60] K. Wada and T. Shibata, "Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 972–980, 2007.

[61] A. A. Scoglio, E. D. Reilly, J. A. Gorman, and C. E. Drebing, "Use of social robots in mental health and well-being research: systematic review," *Journal of Medical Internet Research*, vol. 21, no. 7, p. e13322, 2019.

[62] E. Kim, R. Paul, F. Shic, and B. Scassellati, "Bridging the research gap: Making HRI useful to individuals with autism," 2012.

[63] J. L. Matson and M. J. Konst, "What is the evidence for long term effects of early autism interventions?" *Research in Autism Spectrum Disorders*, vol. 7, no. 3, pp. 475–479, 2013.

[64] S. Jeong, S. Alghowinem, L. Aymerich-Franch, K. Arias, A. Lapedriza, R. Picard, H. W. Park, and C. Breazeal, "A robotic positive psychology coach to improve college students' wellbeing," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 187–194.

[65] K. Jeong, J. Sung, H.-S. Lee, A. Kim, H. Kim, C. Park, Y. Jeong, J. Lee, and J. Kim, "Fribo: A social networking robot for increasing social connectedness through sharing daily home activities from living noise data," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2018, pp. 114–122.

[66] H. R. Pelikan, M. Broth, and L. Keevallik, "'Are you sad, Cozmo?' How humans make sense of a home robot's emotion displays," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 461–470.

[67] H. Chen, A. K. Ostrowski, S. J. Jang, C. Breazeal, and H. W. Park, "Designing long-term parent-child-robot triadic interaction at home through lived technology experiences and interviews," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 401–408.

[68] B. Cagiltay, N. T. White, R. Ibtasar, B. Mutlu, and J. Michaelis, "Understanding factors that shape children's long term engagement with an in-home learning companion robot," in *Interaction Design and Children*, 2022, pp. 362–373.

[69] C. V. Smedegaard, "Novelty knows no boundaries: why a proper investigation of novelty effects within SHRI should begin by addressing the scientific plurality of the field," *Frontiers in Robotics and AI*, vol. 9, p. 741478, 2022.

[70] J. Sung, H. I. Christensen, and R. E. Grinter, "Robots in the wild: understanding long-term use," in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009, pp. 45–52.

[71] C. V. Smedegaard, "Reframing the role of novelty within social hri: from noise to information," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 411–420.

[72] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[73] A. Ramachandran, "Building effective robot tutoring interactions for children," Ph.D. dissertation, Yale University, 2018.

[74] C. Odabasi, F. Graf, J. Lindermayr, M. Patel, S. D. Baumgarten, and B. Graf, "Refilling water bottles in elderly care homes with the help of a safe service robot," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 101–110.

[75] A. Nanavati, N. Walker, L. Taber, C. Mavrogiannis, L. Takayama, M. Cakmak, and S. Srinivasa, "Not all who wander are lost: A localization-free system for in-the-wild mobile robot deployments," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 422–431.

[76] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "A communication robot in a shopping mall," *IEEE Transactions on Robotics*, vol. 26, no. 5, pp. 897–913, 2010.

[77] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz *et al.*, "Designing robots for long-term social interaction," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2005, pp. 1338–1343.

[78] M. Hanheide, D. Hebesberger, and T. Krajník, "The when, where, and how: An adaptive robotic info-terminal for care home residents," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 341–349.

[79] N. Tsoi, J. Connolly, E. Adéníran, A. Hansen, K. T. Pineda, T. Adamson, S. Thompson, R. Ramnauth, M. Vázquez, and B. Scassellati, "Challenges deploying robots during a pandemic: An effort to fight social isolation among children," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021, pp. 234–242.

[80] A. Cesta, G. Cortellessa, A. Orlandini, and L. Tiberio, "Long-term evaluation of a telepresence robot for the elderly: methodology and ecological case study," *International Journal of Social Robotics (IJSR)*, vol. 8, pp. 421–441, 2016.

[81] A. Ahtinen, N. Beheshtian, and K. Väänänen, "Robocamp at home: Exploring families' co-learning with a social robot: Findings from a one-month study in the wild," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2023, pp. 331–340.

[82] M. K. Lee, J. Forlizzi, P. E. Rybski, F. Crabbe, W. Chung, J. Finkle, E. Glaser, and S. Kiesler, "The snackbot: documenting the design of a robot for long-term human-robot interaction," in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009, pp. 7–14.

[83] E. J. Van Der Drift, R.-J. Beun, R. Looije, O. A. Blanson Henkemans, and M. A. Neerincx, "A remote social robot to motivate and support diabetic children in keeping a diary," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2014, pp. 463–470.

[84] K. Severinson-Eklundh, A. Green, and H. Hüttenrauch, "Social and collaborative aspects of interaction with a service robot," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 223–234, 2003.

[85] H. Kozima, M. P. Michalowski, and C. Nakagawa, "Keepon: a playful robot for research, therapy, and entertainment," *International Journal of Social Robotics (IJSR)*, vol. 1, pp. 3–18, 2009.

[86] N. F. Tolksdorf and K. J. Rohlfing, "Parents' views on using social robots for language learning," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 634–640.

[87] D. François, S. Powell, and K. Dautenhahn, "A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children's proactivity and initiative-taking," *Interaction Studies*, vol. 10, no. 3, pp. 324–373, 2009.

[88] A. Vishwanath, A. Singh, Y. H. V. Chua, J. Dauwels, and N. Magnenat-Thalmann, "Humanoid co-workers: How is it like to work with a robot?" in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–6.

[89] K. Stubbs, D. Bernstein, K. Crowley, and I. R. Nourbakhsh, "Long-term human-robot interaction: The personal exploration rover and museum docents." in *AIED*, 2005, pp. 621–628.

[90] E. Gamborino, H.-P. Yueh, W. Lin, S.-L. Yeh, and L.-C. Fu, "Mood estimation as a social profile predictor in an autonomous, multi-session, emotional support robot for children," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–6.

[91] M. Nalin, I. Baroni, I. Kruijff-Korbayová, L. Canamero, M. Lewis, A. Beck, H. Cuayáhuitl, and A. Sanna, "Children's adaptation in multi-session interaction with a humanoid robot," in *The 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2012, pp. 351–357.

[92] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, "Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study," in *The 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2007, pp. 564–569.

[93] A. Coninx, P. Baxter, E. Oleari, S. Bellini, B. Bierman, O. Henkemans, L. Cañamero, P. Cosi, V. Enescu, R. Espinoza *et al.*, "Towards long-term social child-robot interaction: using multi-activity switching to engage young users," *Journal of Human-Robot Interaction*, 2016.

[94] G. W. Lane, D. Noronha, A. Rivera, K. Craig, C. Yee, B. Mills, and E. Villanueva, "Effectiveness of a social robot, 'Paro,' in a VA long-term care setting." *Psychological Services*, vol. 13, no. 3, p. 292, 2016.

[95] G. Nie, Z. Zheng, J. Johnson, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Predicting response to joint attention performance in human-human interaction based on human-robot interaction for young children with autism spectrum disorder," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 1–4.

[96] M. De Graaf, S. Ben Allouch, and J. Van Dijk, "Why do they refuse to use my robot? Reasons for non-use derived from a long-term home study," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 224–233.

[97] M. M. De Graaf, S. Ben Allouch, and J. A. van Dijk, "Long-term evaluation of a social robot in real homes," *Interaction studies*, vol. 17, no. 3, pp. 462–491, 2016.

[98] M. E. Ligthart, S. M. de Droog, M. Bossema, L. Elloumi, K. Hoogland, M. H. Smakman, K. V. Hindriks, and S. Ben Allouch, "Design specifications for a social robot math tutor," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2023, pp. 321–330.

[99] T. Salter, K. Dautenhahn, and R. Bockhorst, "Robots moving out of the laboratory-detecting interaction levels and human contact in noisy school environments," in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*. IEEE, 2004, pp. 563–568.

[100] R. Agrigoroaie and A. Tapus, "The outcome of a week of intensive cognitive stimulation in an elderly care setup: a pilot test," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 814–819.

[101] K. Wada, Y. Takasawa, and T. Shibata, "Robot therapy at facilities for the elderly in Kanagawa prefecture—a report on the experimental result of the first week," in *The 22nd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2013, pp. 757–761.

[102] N. Gasteiger, H. S. Ahn, C. Fok, J. Lim, C. Lee, B. A. MacDonald, G. H. Kim, and E. Broadbent, "Older adults' experiences and perceptions of living with bomy, an assistive dailycare robot: a qualitative study," *Assistive Technology*, vol. 34, no. 4, pp. 487–497, 2022.

[103] T. Klamer, S. Ben Allouch, and D. Heylen, "'adventures of Harvey'–use, acceptance of and relationship building with a social robot in a domestic environment," in *Human-Robot Personal Relationships: Third International Conference, HRPR 2010, Leiden, The Netherlands, June 23-24, 2010, Revised Selected Papers 3*. Springer, 2011, pp. 74–82.

[104] M. Edirisinghe and A. B. P. Jayasekara, "Use of autobiographical memory for enhancing adaptive HRI in multi-user domestic environment," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 1118–1123.

[105] M. I. Ahmad, O. Mubin, and J. Orlando, "Adaptive social robot for sustaining social engagement during long-term children–robot interaction," *International Journal of Human–Computer Interaction*, vol. 33, no. 12, pp. 943–962, 2017.

[106] D. Silvera-Tawil and C. Roberts-Yates, "Socially-assistive robots to enhance learning for secondary students with intellectual disabilities and autism," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 838–843.

[107] B. Robins, K. Dautenhahn, R. T. Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?" *Universal Access in the Information Society*, vol. 4, pp. 105–120, 2005.

[108] F. Tanaka, A. Cicourel, and J. R. Movellan, "Socialization between toddlers and robots at an early childhood education center," *Proceedings of the National Academy of Sciences*, vol. 104, no. 46, pp. 17 954–17 958, 2007.

[109] Y. Fernaeus, M. Håkansson, M. Jacobsson, and S. Ljungblad, "How do you play with a robotic toy animal? a long-term study of Pleo," in *Proceedings of the 9th International Conference on Interaction Design and Children*, 2010, pp. 39–48.

[110] K. Obayashi, N. Kodate, and S. Masuyama, "Assessing the impact of an original soft communicative robot in a nursing home in Japan: Will softness or conversations bring more smiles to older people?" *International Journal of Social Robotics (IJSR)*, vol. 14, no. 3, pp. 645–656, 2022.

[111] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 99–105.

[112] J. B. Janssen, C. C. van der Wal, and M. Neerincx, "Motivate to learn: Effects of performance adaptation on child motivation of robot interaction," 2011.

[113] P. Baxter, E. Ashurst, R. Read, J. Kennedy, and T. Belpaeme, "Robot education peers in a situated primary school study: Personalisation promotes child learning," *PloS One*, vol. 12, no. 5, p. e0178126, 2017.

[114] C. Rivoire and A. Lim, "Habit detection within a long-term interaction with a social robot: an exploratory study," in *Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents*, 2016, pp. 1–6.

[115] J. Piasek and K. Wieczorowska-Tobis, "Acceptance and long-term use of a social robot by elderly users in a domestic environment," in *2018 11th International cConference on Human System Interaction (HSI)*. IEEE, 2018, pp. 478–482.

[116] C. Di Napoli, G. Ercolano, and S. Rossi, "Personalized home-care support for the elderly: a field experience with a social robot at home," *User Modeling and User-Adapted Interaction*, pp. 1–36, 2022.

[117] M. Luperto, J. Monroy, J. Renoux, F. Lunardini, N. Basilico, M. Bulgheroni, A. Cangelosi, M. Cesari, M. Cid, A. Ianes *et al.*, "Integrating social assistive robots, IoT, virtual communities and smart objects to assist at-home independently living elders: the MoveCare project," *International Journal of Social Robotics (IJSR)*, pp. 1–29, 2022.

[118] S. Lemaignan, N. Newbutt, L. Rice, and J. Daly, "'it's important to think of pepper as a teaching aid or resource external to the classroom': A social robot in a school for autistic children," *International Journal of Social Robotics (IJSR)*, pp. 1–22, 2022.

[119] S. Coşar, M. Fernandez-Carmona, R. Agrigoroaie, J. Pages, F. Ferland, F. Zhao, S. Yue, N. Bellotto, and A. Tapus, "Enrichme: Perception and interaction of an assistive robot for the elderly at home," *International Journal of Social Robotics (IJSR)*, vol. 12, pp. 779–805, 2020.

[120] J. E. Michaelis and B. Mutlu, "Reading socially: Transforming the in-home reading experience with a learning-companion robot," *Science Robotics*, vol. 3, no. 21, p. eaat5999, 2018.

[121] S. Šabanović, S. M. Reeder, and B. Kechavarzi, "Designing robots in the wild: In situ prototype evaluation for a break management robot," *Journal of Human-Robot Interaction*, vol. 3, no. 1, pp. 70–88, 2014.

[122] S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, and M. J. Matarić, "Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders," *Science Robotics*, vol. 5, no. 39, p. eaaz3791, 2020.

[123] M. Rueben, J. Klow, M. Duer, E. Zimmerman, J. Piacentini, M. Browning, F. J. Bernieri, C. M. Grimm, and W. D. Smart, "Mental models of a mobile shoe rack: Exploratory findings from a long-term in-the-wild study," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 2, pp. 1–36, 2021.

[124] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, “Interactive robots as social partners and peer tutors for children: A field trial,” *Human–Computer Interaction*, vol. 19, no. 1-2, pp. 61–84, 2004.

[125] I. Leite, C. Martinho, A. Pereira, and A. Paiva, “As time goes by: Long-term evaluation of social presence in robotic companions,” in *The 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2009, pp. 669–674.

[126] N. Salomons, T. Wallenstein, D. Ghose, and B. Scassellati, “The impact of an in-home co-located robotic coach in helping people make fewer exercise mistakes,” in *2022 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN).* IEEE, 2022, pp. 149–154.

[127] A. Ramachandran, C.-M. Huang, and B. Scassellati, “Toward effective robot–child tutoring: Internal motivation, behavioral intervention, and learning outcomes,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 9, no. 1, pp. 1–23, 2019.

[128] J. K. Westlund, G. Gordon, S. Spaulding, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, “Lessons from teachers on performing HRI studies with young children in schools,” in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 2016, pp. 383–390.

[129] A. Jacq, S. Lemaignan, F. Garcia, P. Dillenbourg, and A. Paiva, “Building successful long child-robot interactions in a learning context,” in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 2016, pp. 239–246.

[130] S. Chandra, R. Paradeda, H. Yin, P. Dillenbourg, R. Prada, and A. Paiva, “Do children perceive whether a robotic peer is learning or not?” in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2018, pp. 41–49.

[131] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, “Empathic robots for long-term interaction: evaluating social presence, engagement and perceived support in children,” *International Journal of Social Robotics (IJSR)*, vol. 6, pp. 329–341, 2014.

[132] K. Wada and T. Shibata, “Robot therapy in a care house—its sociopsychological and physiological effects on the residents,” in *Proceedings 2006 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, 2006, pp. 3966–3971.

[133] L. Giusti and P. Marti, “Interpretative dynamics in human robot interaction,” in *The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2006, pp. 111–116.

[134] D. Cruz-Sandoval, A. Morales-Tellez, E. B. Sandoval, and J. Favela, "A social robot as therapy facilitator in interventions to deal with dementia-related behavioral symptoms," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 161–169.

[135] J. Fan, A. Ullal, L. Beuscher, L. C. Mion, P. Newhouse, and N. Sarkar, "Field testing of Ro-Tri, a robot-mediated triadic interaction for older adults," *International Journal of Social Robotics (IJSR)*, vol. 13, pp. 1711–1727, 2021.

[136] N. Gasteiger, H. S. Ahn, C. Gasteiger, C. Lee, J. Lim, C. Fok, B. A. Macdonald, G. H. Kim, and E. Broadbent, "Robot-delivered cognitive stimulation games for older adults: Usability and acceptability evaluation," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 4, pp. 1–18, 2021.

[137] I. Sahin, L. Bashir, A. Arnold, H. G. Tanner, and E. Kokkoni, "Network analysis of an infant's motor actions performed in a robot-assisted learning environment," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 757–762.

[138] K. Wada, Y. Takasawa, and T. Shibata, "Robot therapy at facilities for the elderly in Kanagawa prefecture—a report on the experimental result of the first month," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2014, pp. 193–198.

[139] M. Spitale, M. Axelsson, and H. Gunes, "Robotic mental well-being coaches for the workplace: An in-the-wild study on form," pp. 301–310, 2023.

[140] J. Nakanishi, J. Baba, and H. Ishiguro, "Robot-mediated interaction between children and older adults: A pilot study for greeting tasks in nursery schools," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 63–70.

[141] M. E. Ligthart, M. A. Neerincx, and K. V. Hindriks, "Memory-based personalization for fostering a long-term child-robot relationship," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 80–89.

[142] J. Sung, H. I. Christensen, and R. E. Grinter, "Robots in the wild: understanding long-term use," in *Proceedings of the 4th ACM/IEEE International Conference on Human robot interaction*, 2009, pp. 45–52.

[143] A. M. Sabelli, T. Kanda, and N. Hagita, "A conversational robot in an elderly care center: an ethnographic study," in *Proceedings of the 6th International Conference on Human-Robot Interaction (HRI)*, 2011, pp. 37–44.

[144] F. Michaud, T. Salter, A. Duquette, H. Mercier, M. Lauria, H. Larouche, and F. Larose, "Assistive technologies and child-robot interaction," in *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, 2007.

[145] A. Zhanatkyzy, Z. Telisheva, A. Amirova, N. Rakhymbayeva, and A. Sandygulova, "Multi-purposeful activities for robot-assisted autism therapy: What works best for children's social outcomes?" in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2023, pp. 34–43.

[146] J. Greczek, E. Kaszubski, A. Atrash, and M. Matarić, "Graded cueing feedback in robot-mediated imitation practice for children with autism spectrum disorders," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2014, pp. 561–566.

[147] M. Begum, R. W. Serna, D. Kontak, J. Allspaw, J. Kuczynski, H. A. Yanco, and J. Suarez, "Measuring the efficacy of robots in autism therapy: How informative are standard HRI metrics'," in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 335–342.

[148] N. Rakhymbayeva, A. Amirova, and A. Sandygulova, "A long-term engagement with a social robot for autism therapy," *Frontiers in Robotics and AI*, vol. 8, p. 669972, 2021.

[149] P. Chevalier, G. Raiola, J.-C. Martin, B. Isableu, C. Bazile, and A. Tapus, "Do sensory preferences of children with autism impact an imitation task with a robot?" in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 177–186.

[150] E. Short, K. Swift-Spong, J. Greczek, A. Ramachandran, A. Litoiu, E. C. Grigore, D. Feil-Seifer, S. Shuster, J. J. Lee, S. Huang *et al.*, "How to train your Dragonbot: Socially assistive robots for teaching children about nutrition through play," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2014, pp. 924–929.

[151] L. Süssenbach, N. Riether, S. Schneider, I. Berger, F. Kummert, I. Lütkebohle, and K. Pitsch, "A robot as fitness companion: towards an interactive action-based motivation model," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2014, pp. 286–293.

[152] M. M. De Graaf, S. B. Allouch, and T. Klamer, "Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot," *Computers in Human Behavior*, vol. 43, pp. 1–14, 2015.

[153] R. Feingold Polak and S. L. Tzedek, "Social robot for rehabilitation: expert clinicians and post-stroke patients' evaluation following a long-term intervention," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 151–160.

[154] F. Martí Carrillo, J. Butchart, S. Knight, A. Scheinberg, L. Wise, L. Sterling, and C. McCarthy, "Adapting a general-purpose social robot for paediatric

rehabilitation through in situ design," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 7, no. 1, pp. 1–30, 2018.

[155] D. Hebesberger, T. Koertner, C. Gisinger, J. Pripfl, and C. Dondrup, "Lessons learned from the deployment of a long-term autonomous robot as companion in physical therapy for older adults with dementia a mixed methods study," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 27–34.

[156] L. Taylor, A. Downing, G. A. Noury, G. Masala, M. Palomino, C. McGinn, and R. Jones, "Exploring the applicability of the socially assistive robot Stevie in a day center for people with dementia," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 957–962.

[157] L. McCallum and P. W. McOwan, "Face the music and glance: How nonverbal behaviour aids human robot relationships based in music," in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 237–244.

[158] J. Movellan, M. Eckhardt, M. Virnes, and A. Rodriguez, "Sociable robot improves toddler vocabulary skills," in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009, pp. 307–308.

[159] Z. Shi, T. R. Groechel, S. Jain, K. Chima, O. Rudovic, and M. J. Matarić, "Toward personalized affect-aware socially assistive robot tutors for long-term interventions with children with autism," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 4, pp. 1–28, 2022.

[160] P. Vogt, R. van den Berghe, M. De Haas, L. Hoffman, J. Kanero, E. Mamus, J.-M. Montanier, C. Oranç, O. Oudgenoeg-Paz, D. H. García *et al.*, "Second language tutoring using social robots: a large-scale study," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 497–505.

[161] K. L. Koay, D. S. Syrdal, W. C. Ho, and K. Dautenhahn, "Prototyping realistic long-term human-robot interaction for the study of agent migration," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 809–816.

[162] D. Leyzberg, A. Ramachandran, and B. Scassellati, "The effect of personalization in longer-term robot tutoring," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 7, no. 3, pp. 1–19, 2018.

[163] M. Alemi, A. Meghdari, and M. Ghazisaedy, "Employing humanoid robots for teaching english language in Iranian junior high-schools," *International Journal of Humanoid Robotics*, vol. 11, no. 03, p. 1450022, 2014.

[164] A. Ramachandran, A. Litoiu, and B. Scassellati, "Shaping productive help-seeking behavior during robot-child tutoring interactions," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 247–254.

[165] D. K. Singh, M. Kumar, E. Fosch-Villaronga, D. Singh, and J. Shukla, "Ethical considerations from child-robot interactions in under-resourced communities," *International Journal of Social Robotics (IJSR)*, pp. 1–17, 2022.

[166] X. Zhang, C. Breazeal, and H. W. Park, "A social robot reading partner for explorative guidance," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2023, pp. 341–349.

[167] J. E. Michaelis, B. Cagiltay, R. Ibtasar, and B. Mutlu, "'Off Script:' design opportunities emerging from long-term social robot interactions in-the-wild," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2023, pp. 378–387.

[168] N. Disruption. (2020) Jibo. [Online]. Available: https://jibo.com/

[169] Aldebaran United Robotics Group. (2022) Nao. [Online]. Available: https://www.aldebaran.com/en/nao

[170] BeatBots, "My Keepon," n.d., accessed: 2024-11-05. [Online]. Available: https://www.beatbots.net/my-keepon

[171] Indiegogo, "Jibo: The world's first social robot for the home," https://www.indiegogo.com/projects/jibo-the-world-s-first-social-robot-for-the-home#/, 2014, accessed: 2024-11-05.

[172] PARO Robots U.S., Inc., "PARO therapeutic robot," http://www.parorobots.com/, 2014, accessed: 2024-11-05.

[173] SoftBank Robotics, "Pepper the humanoid robot," n.d., accessed: 2024-11-05. [Online]. Available: https://us.softbankrobotics.com/pepper

[174] A. Steinfeld, O. C. Jenkins, and B. Scassellati, "The oz of wizard: simulating the human for interaction research," in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009, pp. 101–108.

[175] M. Hellou, N. Gasteiger, J. Y. Lim, M. Jang, and H. S. Ahn, "Personalization and localization in human-robot interaction: A review of technical methods," *Robotics*, vol. 10, no. 4, p. 120, 2021.

[176] M. I. Ahmad, O. Mubin, and J. Orlando, "A systematic review of adaptivity in human-robot interaction," *Multimodal Technologies and Interaction*, vol. 1, no. 3, p. 14, 2017.

[177] C. Clabaugh, K. Mahajan, S. Jain, R. Pakkar, D. Becerra, Z. Shi, E. Deng, R. Lee, G. Ragusa, and M. Matarić, "Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders," *Frontiers in Robotics and AI*, vol. 6, p. 110, 2019.

[178] A. Sandygulova, A. Amirova, Z. Telisheva, A. Zhanatkyzy, and N. Rakhymbayeva, "Individual differences of children with autism in robot-assisted autism therapy," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 43–52.

[179] World Health Organization, "Mental health of adolescents," 2021. [Online]. Available: "https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health"

[180] R. L. Katz, M. Felix, and M. Gubernick, "Technology and adolescents: Perspectives on the things to come," *Education and Information Technologies*, vol. 19, pp. 863–886, 2014.

[181] R. E. Grinter and L. Palen, "Instant messaging in teen life," in *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, 2002, pp. 21–30.

[182] D. J. Schiano, C. P. Chen, E. Isaacs, J. Ginsberg, U. Gretarsdottir, and M. Huddleston, "Teen use of messaging media," in *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, 2002, pp. 594–595.

[183] A. R. Lauricella, D. P. Cingel, C. Blackwell, E. Wartella, and A. Conway, "The mobile generation: Youth and adolescent ownership and use of new media," *Communication Research Reports*, vol. 31, no. 4, pp. 357–364, 2014.

[184] House of Commons Library, "Manufacturing: Key economic indicators," 2023. [Online]. Available: https://commonslibrary.parliament.uk/research-briefings/sn05206/

[185] Grand View Research, "Collaborative robots market size, share trends analysis report by payload capacity, by application (assembly, handling, packaging, quality testing), by vertical, by region, and segment forecasts, 2023–2030," 2023. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/collaborative-robots-market

[186] E. Dahlin, "Are robots stealing our jobs?" *Socius*, vol. 5, p. 2378023119846249, 2019.

[187] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics (IJSR)*, vol. 1, pp. 71–81, 2009.

[188] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, “The robotic social attributes scale (rosas) development and validation,” in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 254–262.

[189] T. Nomura, T. Suzuki, T. Kanda, and K. Kato, “Measurement of negative attitudes toward robots,” *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, vol. 7, no. 3, pp. 437–454, 2006.

[190] M. K. Lee, J. Forlizzi, S. Kiesler, P. Rybski, J. Antanitis, and S. Savetsila, “Personalization in HRI: A longitudinal field experiment,” in *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2012, pp. 319–326.

[191] N. Gasteiger, M. Hellou, and H. S. Ahn, “Factors for personalization and localization to optimize human–robot interaction: A literature review,” *International Journal of Social Robotics (IJSR)*, vol. 15, no. 4, pp. 689–701, 2023.

[192] L. H. Waterhouse and L. Waterhouse, *Rethinking autism: Variation and complexity.* Academic Press, 2012.

[193] N. E. Rosen, C. Lord, and F. R. Volkmar, “The diagnosis of autism: From kanner to DSM-III to DSM-5 and beyond,” *Journal of Autism and Developmental Disorders (JADD)*, vol. 51, pp. 4253–4270, 2021.

[194] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders (DSM-5).* American Psychiatric Pub, 2013.

[195] J. J. LeBlanc and M. Fagiolini, “Autism: a ‘critical period’ disorder?” *Neural Plasticity*, vol. 2011, no. 1, p. 921680, 2011.

[196] O. Marín, “Developmental timing and critical windows for the treatment of psychiatric disorders,” *Nature Medicine*, vol. 22, no. 11, pp. 1229–1238, 2016.

[197] R. J. Landa, “Efficacy of early interventions for infants and young children with, and at risk for, autism spectrum disorders,” *International Review of Psychiatry*, vol. 30, no. 1, pp. 25–39, 2018.

[198] K. Bearss, S.-J. Kim, and J. Locke, “Accessible, equitable, and personalized care for autistic individuals,” p. 5217, 2022.

[199] J. Locke, A. Olsen, R. Wideman, M. M. Downey, M. Kretzmann, C. Kasari, and D. S. Mandell, “A tangled web: The challenges of implementing an evidence-based social engagement intervention for children with autism in urban public school settings,” *Behavior Therapy*, vol. 46, no. 1, pp. 54–67, 2015.

[200] L. A. Bilaver and J. Havlicek, “Foster children with autism spectrum disorder: Prevalence, length of stay, and placement patterns,” *Journal of Public Child Welfare*, vol. 7, no. 5, pp. 496–519, 2013.

[201] E. Weir, C. Allison, and S. Baron-Cohen, "Autistic adults have poorer quality healthcare and worse health based on self-report data," *Molecular Autism*, vol. 13, no. 1, p. 23, 2022.

[202] L. A. Redefer and J. F. Goodman, "Brief report: Pet-facilitated therapy with autistic children," *Journal of Autism and Developmental Disorders (JADD)*, vol. 19, pp. 461–467, 1989.

[203] O. Solomon, "What a dog can do: Children with autism and therapy dogs in social interaction," *Ethos*, vol. 38, no. 1, pp. 143–166, 2010.

[204] P. Mesa-Gresa, H. Gil-Gómez, J.-A. Lozano-Quilis, and J.-A. Gil-Gómez, "Effectiveness of virtual reality for children and adolescents with autism spectrum disorder: an evidence-based systematic review," *Sensors*, vol. 18, no. 8, p. 2486, 2018.

[205] S. Ramdoss, W. Machalicek, M. Rispoli, A. Mulloy, R. Lang, and M. O'Reilly, "Computer-based interventions to improve social and emotional skills in individuals with autism spectrum disorders: A systematic review," *Developmental Neurorehabilitation*, vol. 15, no. 2, pp. 119–135, 2012.

[206] J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell, "The clinical use of robots for individuals with autism spectrum disorders: A critical review," *Research in Autism Spectrum Disorders*, vol. 6, no. 1, pp. 249–262, 2012.

[207] I. Werry, K. Dautenhahn, B. Ogden, and W. Harwin, "Can social interaction skills be taught by a social agent? the role of a robotic mediator in autism therapy," in *Cognitive Technology: Instruments of Mind: 4th International Conference, CT 2001 Coventry, UK, August 6–9, 2001 Proceedings.* Springer, 2001, pp. 57–74.

[208] K. Dautenhahn and I. Werry, "A quantitative technique for analysing robot-human interactions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2. IEEE, 2002, pp. 1132–1138.

[209] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, "Effects of repeated exposure to a humanoid robot on children with autism," in *Designing a More Inclusive World.* Springer, 2004, pp. 225–236.

[210] B. Robins, P. Dickerson, P. Stribling, and K. Dautenhahn, "Robot-mediated joint attention in children with autism: A case study in robot-human interaction," *Interaction Studies*, vol. 5, no. 2, pp. 161–198, 2004.

[211] E. Simonoff, A. Pickles, T. Charman, S. Chandler, T. Loucas, and G. Baird, "Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 47, no. 8, pp. 921–929, 2008.

[212] M.-C. Lai, C. Kassee, R. Besney, S. Bonato, L. Hull, W. Mandy, P. Szatmari, and S. H. Ameis, “Prevalence of co-occurring mental health diagnoses in the autism population: a systematic review and meta-analysis,” *The Lancet Psychiatry*, vol. 6, no. 10, pp. 819–829, 2019.

[213] M. Kirkovski, P. G. Enticott, and P. B. Fitzgerald, “A review of the role of female gender in autism spectrum disorders,” *Journal of Autism and Developmental Disorders (JADD)*, vol. 43, no. 11, pp. 2584–2603, 2013.

[214] M.-C. Lai, M. V. Lombardo, B. Auyeung, B. Chakrabarti, and S. Baron-Cohen, “Sex/gender differences and autism: setting the scene for future research,” *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 54, no. 1, pp. 11–24, 2015.

[215] H. Tager-Flusberg and C. Kasari, “Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum,” *Autism Research*, vol. 6, no. 6, pp. 468–478, 2013.

[216] N. Rizvi, W. Wu, M. Bolds, R. Mondal, A. Begel, and I. N. Munyaka, “Are robots ready to deliver autism inclusion?: A critical review,” in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–18.

[217] Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators, “Prevalence of autism spectrum disorders (ASDs)—autism and developmental disabilities monitoring (ADDM) network, United States, 2004,” *MMWR. Surveillance Summaries*, vol. 58, no. SS10, pp. 1–20, 2009. [Online]. Available: https://www.cdc.gov/mmwr/preview/mmwrhtml/ss5810a1.htm

[218] Centers for Disease Control and Prevention, “Data and statistics on autism spectrum disorder,” https://www.cdc.gov/autism/data-research/index.html, May 2025, accessed: 2025-06-09.

[219] B. Scassellati, F. Volkmar, F. Shic, A. Dollar, J. Morrell, and R. Paul, “NSF expeditions in computing: Socially assistive robotics,” https://www.nsf.gov/awardsearch/showAward?AWD_ID=1139078, 2012, national Science Foundation Award #1139078.

[220] J. M. Rehg, M. Clements, G. D. Abowd, and A. Rozga, “Computational behavioral science: Modeling, analysis, and visualization of social and communicative behavior,” https://www.nsf.gov/awardsearch/showAward?AWD_ID=1029679, 2012, national Science Foundation Award #1029679.

[221] “Roboskin: Skin-based technologies and capabilities for safe, autonomous, and interactive robots,” https://cordis.europa.eu/project/id/231500, 2010, eU FP7 ICT Project #231500.

[222] "ALIZ-E: Adaptive strategies for sustainable long-term social interaction," https://cordis.europa.eu/project/id/248116, 2010, eU FP7 ICT Project #248116.

[223] "DREAM: Development of robot-enhanced therapy for children with autism spectrum disorders," https://cordis.europa.eu/project/id/611391, 2015, eU FP7 ICT Project Grant #611391.

[224] B. Scassellati and M. Vázquez, "The potential of socially assistive robots during infectious disease outbreaks," *Science Robotics*, vol. 5, no. 44, 2020.

[225] A. Tapus, M. J. Mataric, and B. Scassellati, "Socially assistive robotics [grand challenges of robotics]," *IEEE robotics & automation magazine*, vol. 14, no. 1, pp. 35–42, 2007.

[226] I. A. Katsanis, V. C. Moulianitis, and D. T. Panagiotarakos, "Design, development, and a pilot study of a low-cost robot for child–robot interaction in autism interventions," *Multimodal Technologies and Interaction*, vol. 6, no. 6, p. 43, 2022.

[227] K. A. Zoder-Martell, A. M. Markelz, M. T. Floress, H. A. Skriba, and L. E. Sayyah, "Technology to facilitate telehealth in applied behavior analysis," *Behavior Analysis in Practice*, vol. 13, pp. 596–603, 2020.

[228] SoftBank Robotics, "NAO robot," https://www.softbankrobotics.com/emea/en/nao, 2022, originally developed by Aldebaran Robotics, now manufactured by SoftBank Robotics.

[229] LuxAI. (n.d.) QTrobot for at-home autism support. Accessed: 4 August 2025. [Online]. Available: https://luxai.com/robot-for-teaching-children-with-autism-at-home/

[230] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[231] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[232] W. Kim, M. Seong, K.-J. Kim, and S. Kim, "Engagnition: A multi-dimensional dataset for engagement recognition of children with autism spectrum disorder," *Scientific Data*, vol. 11, no. 1, p. 299, 2024.

[233] H. Kozima, C. Nakagawa, and Y. Yasuda, "Children–robot interaction: a pilot study in autism therapy," *Progress in Brain Research*, vol. 164, pp. 385–400, 2007.

[234] G. Pioggia, R. Igliozzi, M. L. Sica, M. Ferro, F. Muratori, A. Ahluwalia, D. De Rossi *et al.*, "Exploring emotional and imitational android-based interactions in autistic spectrum disorders," *Journal of CyberTherapy & Rehabilitation*, vol. 1, no. 1, pp. 49–61, 2008.

[235] P. Stribling, J. Rae, and P. Dickerson, "Using conversation analysis to explore the recurrence of a topic in the talk of a boy with an autism spectrum disorder," *Clinical Linguistics & Phonetics*, vol. 23, no. 8, pp. 555–582, 2009.

[236] I. Giannopulu and G. Pradel, "Multimodal interactions in free game play of children with autism and a mobile toy robot," *NeuroRehabilitation*, vol. 27, no. 4, pp. 305–311, 2010.

[237] B. Huskens, R. Verschuur, J. Gillesen, R. Didden, and E. Barakova, "Promoting question-asking in school-aged children with autism spectrum disorders: Effectiveness of a robot intervention compared to a human-trainer intervention," *Developmental Neurorehabilitation*, vol. 16, no. 5, pp. 345–356, 2013.

[238] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati, "Social robots as embedded reinforcers of social behavior in children with autism," *Journal of Autism and Developmental Disorders (JADD)*, vol. 43, pp. 1038–1049, 2013.

[239] O. I. Lovaas, "Behavioral treatment and normal educational and intellectual functioning in young autistic children," *Journal of Consulting and Clinical Psychology*, vol. 55, no. 1, pp. 3–9, 1987.

[240] G. Dawson, S. Rogers, J. Munson, M. Smith, J. Winter, J. Greenson, A. Donaldson, and J. Varley, "Early behavioral intervention is associated with normalized brain activity in young children with autism," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 49, no. 11, pp. 1150–1158, 2010.

[241] R. Paul, "Interventions to improve communication in autism," *Child and Adolescent Psychiatric Clinics of North America*, vol. 17, no. 4, pp. 835–856, 2008.

[242] A. Arnold, R. J. Semple, I. Beale, and C. M. Fletcher-Flinn, "Eye contact in children's social interactions: What is normal behaviour?" *Journal of Intellectual and Developmental Disability*, vol. 25, no. 3, pp. 207–216, 2000.

[243] N. Peters-Scheffer, R. Didden, H. Korzilius, and P. Sturmey, "A meta-analytic study on the effectiveness of comprehensive aba-based early intervention programs for children with autism spectrum disorders," *Research in Autism Spectrum Disorders*, vol. 5, no. 1, pp. 60–69, 2011.

[244] R. M. Foxx, "Applied behavior analysis treatment of autism: The state of the art," *Child and Adolescent Psychiatric Clinics of North America*, vol. 17, no. 4, pp. 821–834, 2008.

[245] K. M. Hustyi, A. H. Ryan, and S. S. Hall, "A scoping review of behavioral interventions for promoting social gaze in individuals with autism spectrum disorder and other developmental disabilities," *Research in Autism Spectrum Disorders*, vol. 100, p. 102074, 2023.

[246] I. Alkinj, A. Pereira, and P. C. Santos, "The effects of an educational program based on modeling and social stories on improvements in the social skills of students with autism," *Heliyon*, vol. 8, no. 5, 2022.

[247] A. Bandura, "Social learning through imitation." 1962.

[248] I. M. Smith and S. E. Bryson, "Imitation and action in autism: A critical review," *Psychological Bulletin*, vol. 116, no. 2, pp. 259–273, 1994.

[249] D. Spain, J. Sin, K. B. Linder, J. McMahon, and F. Happé, "Social anxiety in autism spectrum disorder: A systematic review," *Research in Autism Spectrum Disorders*, vol. 52, pp. 51–68, 2018.

[250] J. B. Leaf, R. Leaf, J. McEachin, M. Taubman, S. Ala'i-Rosales, R. K. Ross, T. Smith, and M. J. Weiss, "Applied behavior analysis is a science and, therefore, progressive," *Journal of Autism and Developmental Disorders (JADD)*, vol. 46, no. 2, pp. 720–731, 2016.

[251] K. Bottema-Beutel, S. Crowley, M. Sandbank, and T. G. Woynaroski, "Research review: Conflicts of interest (COIs) in autism early intervention research–a meta-analysis of coi influences on intervention effects," *Journal of Child Psychology and Psychiatry*, vol. 62, no. 1, pp. 5–15, 2021.

[252] C. A. DiSalvo and D. P. Oswald, "Peer-mediated interventions to increase the social interaction of children with autism: Consideration of peer expectancies," *Focus on autism and other developmental disabilities*, vol. 17, no. 4, pp. 198–207, 2002.

[253] Y.-C. Chang and J. Locke, "A systematic review of peer-mediated interventions for children with autism spectrum disorder," *Research in Autism Spectrum Disorders*, vol. 27, pp. 1–10, 2016.

[254] D. Feil-Seifer and M. J. Matarić, "Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders," in *Experimental Robotics: The Eleventh International Symposium*. Springer, 2009, pp. 201–210.

[255] G. Pioggia, M. Sica, M. Ferro, R. Igliozzi, F. Muratori, A. Ahluwalia, and D. De Rossi, "Human-robot interaction in autism: FACE, an android-based social therapy," in *The 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2007, pp. 605–612.

[256] Y. Nakadoi, "Usefulness of animal type robot assisted therapy for autism spectrum disorder in the child and adolescent psychiatric ward," in *JSAI International Symposium on Artificial Intelligence*. Springer, 2015, pp. 478–482.

[257] S. da Luz Schreider, J. C. de Souza, É. V. da Silva Freitas, J. A. C. Panceri, E. M. de Oliveira Caldeira, and T. F. Bastos-Filho, "Psychomotor intervention through serious games in children and adolescents with autism spectrum disorder using a therapeutic robot," *Research on Biomedical Engineering*, vol. 40, no. 3, pp. 485–497, 2024.

[258] P. Mundy and L. Newell, "Attention, joint attention, and social cognition," *Current Directions in pPsychological Science*, vol. 16, no. 5, pp. 269–274, 2007.

[259] M. Tomasello, "Joint attention as social cognition," in *Joint attention*. Psychology Press, 2014, pp. 103–130.

[260] S. Baron-Cohen, "Theory of mind and autism: A review," *International Review of Research in Mental Retardation*, vol. 23, pp. 169–184, 2000.

[261] J. L. Bean and I.-M. Eigsti, "Assessment of joint attention in school-age children and adolescents," *Research in Autism Spectrum Disorders*, vol. 6, no. 4, pp. 1304–1310, 2012.

[262] A. Taheri, A. Meghdari, and M. H. Mahoor, "A close look at the imitation performance of children with autism and typically developing children using a robotic system," *International Journal of Social Robotics (IJSR)*, vol. 13, no. 5, pp. 1125–1147, 2021.

[263] A. Zhanatkyzy, Z. Telisheva, A. Amirova, N. Rakhymbayeva, and A. Sandygulova, "Multi-purposeful activities for robot-assisted autism therapy: What works best for children's social outcomes?" in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2023, pp. 34–43.

[264] S. M. Srinivasan, K. A. Lynch, D. J. Bubela, T. D. Gifford, and A. N. Bhat, "Effect of interactions between a child and a robot on the imitation and praxis performance of typically devloping children and a child with autism: A preliminary study," *Perceptual and Motor Skills*, vol. 116, no. 3, pp. 885–904, 2013.

[265] S. Costa, C. Santos, F. Soares, M. Ferreira, and F. Moreira, "Promoting interaction amongst autistic adolescents using robots," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 3856–3859.

[266] A. Duquette, F. Michaud, and H. Mercier, "Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism," *Autonomous Robots*, vol. 24, no. 2, pp. 147–157, 2008.

[267] B. Robins, K. Dautenhahn, R. T. Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?" *Universal Access in the Information Society*, vol. 4, no. 2, pp. 105–120, 2005.

[268] S. Shamsuddin, H. Yussof, L. Ismail, F. A. Hanapiah, S. Mohamed, H. A. Piah, and N. I. Zahari, "Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO," in *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*. IEEE, 2012, pp. 188–193.

[269] I. Ranatunga, N. A. Torres, R. Patterson, N. Bugnariu, M. Stevenson, and D. O. Popa, "Rodica: a human-robot interaction system for treatment of childhood autism spectrum disorders," in *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments*, 2012, pp. 1–6.

[270] H. Kozima, M. P. Michalowski, and C. Nakagawa, "Keepon: A playful robot for research, therapy, and entertainment," *International Journal of Social Robotics (IJSR)*, vol. 1, no. 1, pp. 3–18, 2009.

[271] B. Robins, K. Dautenhahn, and P. Dickerson, "From isolation to communication: a case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot," in *2009 Second International Conferences on Advances in Computer-Human Interactions*. IEEE, 2009, pp. 205–211.

[272] D. Mazzei, N. Lazzeri, L. Billeci, R. Igliozzi, A. Mancini, A. Ahluwalia, F. Muratori, and D. De Rossi, "Development and evaluation of a social robot platform for therapy in autism," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 4515–4518.

[273] M. Gaitán-Padilla, C. A. Cifuentes, and M. Múnera, "Three study cases of social robotics in autism spectrum disorder treatment: Personalization and usability of castor robot," in *2022 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2022, pp. 1–6.

[274] C. A. Huijnen, H. A. Verreussel-Willen, M. A. Lexis, and L. P. de Witte, "Robot KASPAR as mediator in making contact with children with autism: A pilot study," *International Journal of Social Robotics (IJSR)*, vol. 13, no. 2, pp. 237–249, 2021.

[275] B. Robins and K. Dautenhahn, "Tactile interactions with a humanoid robot: novel play scenario implementations with children with autism," *International Journal of Social Robotics (IJSR)*, vol. 6, no. 3, pp. 397–415, 2014.

[276] K. Marathaki, M. Feidakis, C. Patrikakis, and E. Agrianiti, "Deploy social assistive robot to develop symbolic play and imitation skills in students with autism

spectrum disorder," in *2022 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2022, pp. 547–552.

[277] B. Huskens, A. Palmen, M. Van der Werff, T. Lourens, and E. Barakova, "Improving collaborative play between children with autism spectrum disorders and their siblings: The effectiveness of a robot-mediated intervention based on Lego therapy," *Journal of Autism and Developmental Disorders (JADD)*, vol. 45, no. 11, pp. 3746–3755, 2015.

[278] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn, "Using the humanoid robot KASPAR to autonomously play triadic games and facilitate collaborative play among children with autism," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 3, pp. 183–199, 2014.

[279] M. Dimitrova, N. Vegt, and E. Barakova, "Designing a system of interactive robots for training collaborative skills to autistic children," in *2012 15th International Conference on Interactive Collaborative Learning (ICL)*. IEEE, 2012, pp. 1–8.

[280] S. Pliasa and N. Fachantidis, "Can a robot be an efficient mediator in promoting dyadic activities among children with autism spectrum disorders and children of typical development?" in *Proceedings of the 9th Balkan Conference on Informatics*, 2019, pp. 1–6.

[281] F. S. Melo, A. Sardinha, D. Belo, M. Couto, M. Faria, A. Farias, H. Gamboa, C. Jesus, M. Kinarullathil, P. Lima *et al.*, "Project INSIDE: towards autonomous semi-unstructured human–robot social interaction in autism therapy," *Artificial Intelligence in Medicine*, vol. 96, pp. 198–216, 2019.

[282] J. Lee, H. Takehashi, C. Nagai, G. Obinata, and D. Stefanov, "Which robot features can stimulate better responses from children with autism in robot-assisted therapy?" *International Journal of Advanced Robotic Systems*, vol. 9, no. 3, p. 72, 2012.

[283] K. Conn, C. Liu, N. Sarkar, W. Stone, and Z. Warren, "Affect-sensitive assistive intervention technologies for children with autism: An individual-specific approach," in *The 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2008, pp. 442–447.

[284] S. Lemaignan, N. Newbutt, L. Rice, and J. Daly, "'it's important to think of Pepper as a teaching aid or resource external to the classroom': A social robot in a school for autistic children," *International Journal of Social Robotics (IJSR)*, vol. 16, no. 6, pp. 1083–1104, 2024.

[285] B. Xie and C. H. Park, "An empathetic social robot with modular anxiety interventions for autistic adolescents," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2024, pp. 1148–1155.

[286] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, B. A. Corbett, Y. Matsumoto, H. Higashida, T. Yuhi, H. Ishiguro, M. Mimura, and M. Kikuchi, "Job interview training targeting nonverbal communication using an android robot for individuals with autism spectrum disorder," *Autism*, vol. 23, no. 6, pp. 1586–1595, 2019.

[287] P. Shahverdi, K. Rousso, I. Bakhoda, N. Huang, K. Rohrbeck, and W.-Y. G. Louie, "Robot-mediated job interview training for individuals with ASD: a pilot study," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 564–570.

[288] H. Kumazaki, Z. Warren, B. A. Corbett, Y. Yoshikawa, Y. Matsumoto, H. Higashida, T. Yuhi, T. Ikeda, H. Ishiguro, and M. Kikuchi, "Android robot-mediated mock job interview sessions for young adults with autism spectrum disorder: A pilot study," *Frontiers in Psychiatry*, vol. 8, p. 169, 2017.

[289] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, Y. Matsumoto, M. Miyao, H. Ishiguro, M. Mimura, Y. Minabe, and M. Kikuchi, "How the realism of robot is needed for individuals with autism spectrum disorders in an interview setting," *Frontiers in Psychiatry*, vol. 10, p. 486, 2019.

[290] G. Vivanti, "Ask the editor: What is the most appropriate way to talk about individuals with a diagnosis of autism?" *Journal of Autism and Developmental Disorders (JADD)*, vol. 50, no. 2, pp. 691–693, 2020.

[291] S. H. Kim and C. Lord, "Autism diagnostic observation schedule," in *Encyclopedia of Clinical Neuropsychology*. Springer, 2018, pp. 1–6.

[292] A. Snow, "Social communication questionnaire," in *Encyclopedia of Autism Spectrum Disorders*. Springer, 2021, pp. 4427–4429.

[293] F. Abu-Amara, H. Mohammad, and A. Bensefia, "Robot-based therapy for improving academic skills of children with autism," *International Journal of Information Technology*, vol. 16, no. 6, pp. 3371–3380, 2024.

[294] I. Giannopulu and G. Pradel, "From child-robot interaction to child-robot-therapist interaction: A case study in autism," *Applied Bionics and Biomechanics*, vol. 9, no. 2, pp. 173–179, 2012.

[295] C. Clabaugh, K. Mahajan, S. Jain, R. Pakkar, D. Becerra, Z. Shi, E. Deng, R. Lee, G. Ragusa, and M. Matarić, "Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders," *Frontiers in Robotics and AI*, vol. 6, p. 110, 2019.

[296] E. Y.-h. Chung, "Robot-mediated social skill intervention programme for children with autism spectrum disorder: An ABA time-series study," *International Journal of Social Robotics (IJSR)*, vol. 13, no. 5, pp. 1095–1107, 2021.

[297] K. Silva, M. Lima, A. Santos-Magalhães, C. Fafiães, and L. de Sousa, "Living and robotic dogs as elicitors of social communication behavior and regulated emotional responding in individuals with autism and severe language delay: A preliminary comparative study," *Anthrozoös*, vol. 32, no. 1, pp. 23–33, 2019.

[298] S. R. McConnell, "Interventions to facilitate social interaction for young children with autism: Review of available research and recommendations for educational intervention and future research," *Journal of Autism and Developmental Disorders (JADD)*, vol. 32, no. 5, pp. 351–372, 2002.

[299] P. Soleiman, H. Moradi, B. Mehralizadeh, H. Ameri, R. I. Arriaga, H. R. Pouretemad, N. Baghbanzadeh, and L. K. Vahid, "Fully robotic social environment for teaching and practicing affective interaction: Case of teaching emotion recognition skills to children with autism spectrum disorder, a pilot study," *Frontiers in Robotics and AI*, vol. 10, p. 1088582, 2023.

[300] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, Y. Yoshimura, T. Ikeda, C. Hasegawa, D. N. Saito, J. Shimaya, H. Ishiguro, M. Mimura *et al.*, "Brief report: A novel system to evaluate autism spectrum disorders using two humanoid robots," *Journal of Autism and Developmental Disorders (JADD)*, vol. 49, no. 4, pp. 1709–1716, 2019.

[301] K. Takata, Y. Yoshikawa, T. Muramatsu, Y. Matsumoto, H. Ishiguro, M. Mimura, and H. Kumazaki, "Social skills training using multiple humanoid robots for individuals with autism spectrum conditions," *Frontiers in Psychiatry*, vol. 14, p. 1168837, 2023.

[302] V. Kostrubiec, C. Lajunta, P.-V. Paubel, and J. Kruck, "Does the social robot Nao facilitate cooperation in high functioning children with ASD?" *International Journal of Social Robotics (IJSR)*, vol. 16, no. 2, pp. 281–298, 2024.

[303] M. Begum, R. W. Serna, and H. A. Yanco, "Are robots ready to deliver autism interventions? a comprehensive review," *International Journal of Social Robotics (IJSR)*, vol. 8, pp. 157–181, 2016.

[304] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, and M. Blow, "KASPAR–a minimally expressive humanoid robot for human–robot interaction research," *Applied Bionics and Biomechanics*, vol. 6, no. 3-4, pp. 369–397, 2009.

[305] P. Soleiman, S. Salehi, M. Mahmoudi, M. Ghavami, H. Moradi, and H. Pouretemad, "RoboParrot: A robotic platform for human robot interaction, case of autistic children," in *2014 Second RSI/ISM International Conference on Robotics and Mechatronics (ICRoM)*. IEEE, 2014, pp. 711–716.

[306] N. Rakhymbayeva, A. Amirova, and A. Sandygulova, "A long-term engagement with a social robot for autism therapy," *Frontiers in Robotics and AI*, vol. 8, p. 669972, 2021.

[307] E. Zehnder, M. Jouaiti, and F. Charpillet, "Evaluating robot acceptance in children with ASD and their parents," in *International Conference on Social Robotics*. Cham: Springer Nature Switzerland, Dec. 2022, pp. 45–53.

[308] I. van den Berk-Smeekens, M. van Dongen-Boomsma, M. De Korte, J. Den Boer, I. Oosterling, N. Peters-Scheffer, J. Buitelaar, E. Barakova, T. Lourens, W. Staal, and J. Glennon, "Adherence and acceptability of a robot-assisted pivotal response treatment protocol for children with autism spectrum disorder," *Scientific Reports*, vol. 10, no. 1, p. 8110, 2020.

[309] J. Frenkel, S. Kirst, S. Naumann, M. Simon, J. Sessner, E. Roesler, L. Onnasch, and I. Dziobek, "Stakeholder acceptance of a robot-assisted social training scenario for autistic children compared to a tablet-computer-based approach," *Scientific Reports*, vol. 15, no. 1, p. 11237, 2025.

[310] A. Sorrentino, G. Mancioppi, L. Coviello, F. Cavallo, and L. Fiorini, "Feasibility study on the role of personality, emotion, and engagement in socially assistive robotics: A cognitive assessment scenario," *Informatics*, vol. 8, no. 2, p. 23, Mar. 2021.

[311] Z. Shi, T. R. Groechel, S. Jain, K. Chima, O. Rudovic, and M. J. Matarić, "Toward personalized affect-aware socially assistive robot tutors for long-term interventions with children with autism," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 4, pp. 1–28, 2022.

[312] Z. Zheng, H. Zhao, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Design, development, and evaluation of a noninvasive autonomous robot-mediated joint attention intervention system for young children with ASD," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 125–135, 2017.

[313] M. Coeckelbergh, C. Pop, R. Simut, A. Peca, S. Pintea, D. David, and B. Vanderborght, "A survey of expectations about the role of robots in robot-assisted therapy for children with ASD: ethical acceptability, trust, sociability, appearance, and attachment," *Science and Engineering Ethics*, vol. 22, pp. 47–65, 2016.

[314] M. Yoshikawa, Y. Matsumoto, M. Sumitani, and H. Ishiguro, "Development of an android robot for psychological support in medical and welfare fields," in *2011 IEEE International Conference on Robotics and Biomimetics*. IEEE, 2011, pp. 2378–2383.

[315] Robokind, *Milo: The Facially Expressive Robot for Autism Intervention.* Robokind, 2018. [Online]. Available: https://www.robokind.com

[316] Osaka University & Vstone Co., Ltd. / Ishiguro Symbiotic HRI Project. (2015) Development of social dialogue robots CommU and Sota. Accessed: 4 August

2025; details on design and launch of CommU and Sota tabletop robots. [Online]. Available: https://resou.osaka-u.ac.jp/en/research/2015/20150120_2

[317] Y. Fernaeus, M. Håkansson, M. Jacobsson, and S. Ljungblad, "How do you play with a robotic toy animal? a long-term study of Pleo," in *Proceedings of the 9th International Conference on Interaction Design and Children*, 2010, pp. 39–48.

[318] H. Kozima, M. P. Michalowski, and C. Nakagawa, "Keepon: A playful robot for research, therapy, and entertainment," *International Journal of Social Robotics (IJSR)*, vol. 1, pp. 3–18, 2009.

[319] A. Billard, B. Robins, J. Nadel, and K. Dautenhahn, "Building Robota, a mini-humanoid robot for the rehabilitation of children with autism," *Assistive Technology*, vol. 19, no. 1, pp. 37–49, 2007.

[320] S. S. Balasuriya, L. Sitbon, M. Brereton, and S. Koplick, "How can social robots spark collaboration and engagement among people with intellectual disability?" in *Proceedings of the 31st Australian Conference on Human-Computer Interaction*, 2019, pp. 209–220.

[321] Soft Interaction Lab, "Soft robot for non-verbal children with autism," https://softinteraction.com/portfolio/1064-2, 2014, accessed: 2025-07-16.

[322] W. D. Stiehl, C. Breazeal, K.-H. Han, J. Lieberman, L. Lalla, A. Maymin, J. Salinas, D. Fuentes, R. Toscano, C. H. Tong *et al.*, "The Huggable: a therapeutic robotic companion for relational, affective touch," in *ACM SIGGRAPH 2006 Emerging Technologies*, 2006, pp. 15–es.

[323] E. Y.-h. Chung, K. K.-f. Sin, and D. H.-k. Chow, "Effectiveness of robotic intervention on improving social development and participation of children with autism spectrum disorder–a randomised controlled trial," *Journal of Autism and Developmental Disorders (JADD)*, vol. 55, no. 2, pp. 449–456, Feb. 2025.

[324] R. Mishra and K. Conn Welch, "Towards scalable robotic intervention of children with autism spectrum disorder using LLMs," *arXiv e-prints*, pp. arXiv–2402, 2024.

[325] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Human–Computer Interaction*, vol. 19, no. 1-2, pp. 61–84, 2004.

[326] S. Shamsuddin, H. Yussof, L. I. Ismail, S. Mohamed, F. A. Hanapiah, and N. I. Zahari, "Initial response in HRI-a case study on evaluation of child with autism spectrum disorders interacting with a humanoid robot Nao," *Procedia Engineering*, vol. 41, pp. 1448–1455, 2012.

[327] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati, "Social robots as embedded reinforcers of social behavior in children with autism," *Journal of Autism and Developmental Disorders (JADD)*, vol. 43, no. 5, pp. 1038–1049, 2013.

[328] A. Zaraki, M. Khamassi, L. J. Wood, G. Lakatos, C. Tzafestas, F. Amirabdollahian, B. Robins, and K. Dautenhahn, "A novel reinforcement-based paradigm for children to teach the humanoid Kaspar robot," *International Journal of Social Robotics (IJSR)*, vol. 12, no. 3, pp. 709–720, 2020.

[329] J. A. Barnes, C. H. Park, A. Howard, and M. Jeon, "Child-robot interaction in a musical dance game: An exploratory comparison study between typically developing children and children with autism," *International Journal of Human–Computer Interaction*, vol. 37, no. 3, pp. 249–266, 2021.

[330] H. Erel, M. Vázquez, S. Sebo, N. Salomons, S. Gillet, and B. Scassellati, "RoSI: A model for predicting robot social influence," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 13, no. 2, pp. 1–22, 2024.

[331] C. V. Smedegaard, "Reframing the role of novelty within social hri: from noise to information," in *2019 14th acm/IEEE International Conference on Human-Robot Interaction (hri)*. IEEE, 2019, pp. 411–420.

[332] M. Begum, R. W. Serna, and H. A. Yanco, "Are robots ready to deliver autism interventions? a comprehensive review," *International Journal of Social Robotics (IJSR)*, vol. 8, pp. 157–181, 2016.

[333] I. van den Berk-Smeekens, M. W. de Korte, M. van Dongen-Boomsma, I. J. Oosterling, J. C. den Boer, E. I. Barakova, T. Lourens, J. C. Glennon, W. G. Staal, and J. K. Buitelaar, "Pivotal response treatment with and without robot-assistance for children with autism: a randomized controlled trial," *European Child & Adolescent Psychiatry*, vol. 31, no. 12, pp. 1871–1883, 2022.

[334] Z. Zheng, G. Nie, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "A randomized controlled trial of an intelligent robotic response to joint attention intervention system," *Journal of Autism and Developmental Disorders (JADD)*, vol. 50, no. 8, pp. 2819–2831, 2020.

[335] F. Marino, P. Chilà, S. T. Sfrazzetto, C. Carrozza, I. Crimi, C. Failla, M. Busà, G. Bernava, G. Tartarisco, D. Vagni *et al.*, "Outcomes of a robot-assisted social-emotional understanding intervention for young children with autism spectrum disorders," *Journal of Autism and Developmental Disorders (JADD)*, vol. 50, no. 6, pp. 1973–1987, 2020.

[336] J. Lee and T. Nagae, "Social distance in interactions between children with autism and robots," *Applied Sciences*, vol. 11, no. 22, p. 10520, 2021.

[337] H. Feng, M. H. Mahoor, and F. Dino, "A music-therapy robotic platform for children with autism: A pilot study," *Frontiers in Robotics and AI*, vol. 9, p. 855819, 2022.

[338] W.-C. So, C.-H. Cheng, W.-Y. Lam, Y. Huang, K.-C. Ng, H.-C. Tung, and W. Wong, "A robot-based play-drama intervention may improve the joint attention and functional play behaviors of chinese-speaking preschoolers with autism spectrum disorder: a pilot study," *Journal of Autism and Developmental Disorders (JADD)*, vol. 50, no. 2, pp. 467–481, 2020.

[339] S.-S. Yun, J. Choi, S.-K. Park, G.-Y. Bong, and H. Yoo, "Social skills training for children with autism spectrum disorder using a robotic behavioral intervention system," *Autism Research*, vol. 10, no. 7, pp. 1306–1323, 2017.

[340] H.-L. Cao, P. G. Esteban, M. Bartlett, P. Baxter, T. Belpaeme, E. Billing, H. Cai, M. Coeckelbergh, C. Costescu, D. David *et al.*, "Robot-enhanced therapy: Development and validation of supervised autonomous robotic system for autism spectrum disorders therapy," *IEEE Robotics & Automation Magazine*, vol. 26, no. 2, pp. 49–58, 2019.

[341] E. Billing, T. Belpaeme, H. Cai, H.-L. Cao, A. Ciocan, C. Costescu, D. David, R. Homewood, D. Hernandez Garcia, P. Gómez Esteban *et al.*, "The DREAM dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy," *PloS One*, vol. 15, no. 8, p. e0236939, 2020.

[342] K. T. Sterrett, *Sensitivity and Expected Change of Commonly Used Social Communication Measures in Longitudinal Research of Children with Autism.* University of California, Los Angeles, 2021.

[343] S. H. Kim, R. Grzadzinski, K. Martinez, and C. Lord, "Measuring treatment response in children with autism spectrum disorder: Applications of the brief observation of social communication change to the autism diagnostic observation schedule," *Autism*, vol. 23, no. 5, pp. 1176–1185, 2019.

[344] S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, and M. J. Matarić, "Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders," *Science Robotics*, vol. 5, no. 39, p. eaaz3791, 2020.

[345] M. Trombly, P. Shahverdi, N. Huang, Q. Chen, J. Korneder, and W.-Y. G. Louie, "Robot-mediated group instruction for children with ASD: a pilot study," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2022, pp. 1506–1513.

[346] A. Alabdulkareem, N. Alhakbani, and A. Al-Nafjan, "A systematic review of research on robot-assisted therapy for children with autism," *Sensors*, vol. 22, no. 3, p. 944, 2022.

[347] J.-J. Cabibihan, H. Javed, M. Ang, and S. M. Aljunied, "Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism," *International Journal of Social Robotics (IJSR)*, vol. 5, pp. 593–618, 2013.

[348] K. Dautenhahn and I. Werry, "Towards interactive robots in autism therapy: Background, motivation and challenges," *Pragmatics & Cognition*, vol. 12, no. 1, pp. 1–35, 2004.

[349] L. A. Dickstein-Fischer, D. E. Crone-Todd, I. M. Chapman, A. T. Fathima, and G. S. Fischer, "Socially assistive robots: current status and future prospects for autism interventions," *Innovation and Entrepreneurship in Health*, pp. 15–25, 2018.

[350] M. A. Islam, M. M. Hasan, and S. A. Deowan, "Robot-assisted training for children with autism spectrum disorder: a review," *Journal of Intelligent & Robotic Systems*, vol. 108, no. 3, p. 41, 2023.

[351] L. I. Ismail, T. Verhoeven, J. Dambre, and F. Wyffels, "Leveraging robotics research for children with autism: a review," *International Journal of Social Robotics (IJSR)*, vol. 11, pp. 389–410, 2019.

[352] A. Kouroupa, K. R. Laws, K. Irvine, S. E. Mengoni, A. Baird, and S. Sharma, "The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis," *PloS One*, vol. 17, no. 6, p. e0269800, 2022.

[353] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, Y. Matsumoto, H. Ishiguro, M. Kikuchi, T. Sumiyoshi, and M. Mimura, "Optimal robot for intervention for individuals with autism spectrum disorders," *Psychiatry and clinical neurosciences*, vol. 74, no. 11, pp. 581–586, 2020.

[354] A. Puglisi, T. Caprì, L. Pignolo, S. Gismondo, P. Chilà, R. Minutoli, F. Marino, C. Failla, A. A. Arnao, G. Tartarisco *et al.*, "Social humanoid robots for children with autism spectrum disorders: a review of modalities, indications, and pitfalls," *Children*, vol. 9, no. 7, p. 953, 2022.

[355] M. A. Saleh, F. A. Hanapiah, and H. Hashim, "Robot applications for autism: a comprehensive review," *Disability and Rehabilitation: Assistive Technology*, vol. 16, no. 6, pp. 580–602, 2021.

[356] B. Szymona, M. Maciejewski, R. Karpiński, K. Jonak, E. Radzikowska-Büchner, K. Niderla, and A. Prokopiak, "Robot-assisted autism therapy (RAAT). criteria and types of experiments using anthropomorphic and zoomorphic robots. review of the research," *Sensors*, vol. 21, no. 11, p. 3720, 2021.

[357] H. Javed and C. H. Park, "Interactions with an empathetic agent: Regulating emotions and improving engagement in autism," *IEEE robotics & automation magazine*, vol. 26, no. 2, pp. 40–48, 2019.

[358] F. Tanaka, A. Cicourel, and J. R. Movellan, "Socialization between toddlers and robots at an early childhood education center," *Proceedings of the National Academy of Sciences*, vol. 104, no. 46, pp. 17 954–17 958, 2007.

[359] P. T. Shattuck, M. Wagner, S. Narendorf, P. Sterzing, and M. Hensley, "Post–high school service use among young adults with an autism spectrum disorder," *Archives of Pediatrics & Adolescent Medicine*, vol. 165, no. 2, pp. 141–146, 2011.

[360] N. A. Henninger and J. L. Taylor, "Outcomes in adults with autism spectrum disorders: A historical perspective," *Autism*, vol. 17, no. 1, pp. 103–116, 2013.

[361] J. E. Graetz, "Autism grows up: Opportunities for adults with autism," *Disability & Society*, vol. 25, no. 1, pp. 33–47, 2010.

[362] E. Müller, A. Schuler, B. A. Burton, and G. B. Yates, "Meeting the vocational support needs of individuals with asperger syndrome and other autism spectrum disabilities," *Journal of Vocational Rehabilitation*, vol. 18, no. 3, pp. 163–175, 2003.

[363] K. A. Anderson, C. Sosnowy, A. A. Kuo, and P. T. Shattuck, "Transition of individuals with autism to adulthood: A review of qualitative studies," *Pediatrics*, vol. 141, no. Supplement_4, pp. S318–S327, 2018.

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

[365] A. Kerepesi, E. Kubinyi, G. K. Jonsson, M. S. Magnússon, and Á. Miklósi, "Behavioural comparison of human–animal (dog) and human–robot (aibo) interactions," *Behavioural Processes*, vol. 73, no. 1, pp. 92–99, 2006.

[366] O. Barber, E. Somogyi, E. A. McBride, and L. Proops, "Exploring the role of aliveness in children's responses to a dog, biomimetic robot, and toy dog," *Computers in Human Behavior*, vol. 142, p. 107660, 2023.

[367] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents," *International Journal of Social Robotics (IJSR)*, vol. 3, pp. 41–52, 2011.

[368] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The physical presence of a robot tutor increases cognitive learning gains," in *Proceedings of the annual meeting of the cognitive science society*, vol. 34, no. 34, 2012.

[369] J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric, "The role of physical embodiment in human-robot interaction," in *The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2006, pp. 117–122.

[370] E. Deng, B. Mutlu, M. J. Mataric *et al.*, "Embodiment in socially interactive robots," *Foundations and Trends in Robotics*, vol. 7, no. 4, pp. 251–356, 2019.

[371] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind"?" *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[372] T. Charman, J. Swettenham, S. Baron-Cohen, A. Cox, G. Baird, and A. Drew, "Infants with autism: an investigation of empathy, pretend play, joint attention, and imitation." *Developmental Psychology*, vol. 33, no. 5, p. 781, 1997.

[373] G. Odekerken-Schröder, C. Mele, T. Russo-Spena, D. Mahr, and A. Ruggiero, "Mitigating loneliness with companion robots in the COVID-19 pandemic and beyond: an integrative framework and research agenda," *Journal of Service Management*, 2020.

[374] J. Holt-Lunstad, "The potential public health relevance of social isolation and loneliness: Prevalence, epidemiology, and risk factors," *Public Policy & Aging Report*, vol. 27, no. 4, pp. 127–130, 2017.

[375] E. Golberstein, H. Wen, and B. F. Miller, "Coronavirus Disease 2019 (COVID-19) and Mental Health for Children and Adolescents," *JAMA Pediatrics*, vol. 174, no. 9, pp. 819–820, 09 2020. [Online]. Available: https://doi.org/10.1001/jamapediatrics.2020.1456

[376] S. A. Lynch and C. G. Simpson, "Social skills: Laying the foundation for success." *Dimensions of Early Childhood*, vol. 38, no. 2, pp. 3–12, 2010.

[377] D. F. Bjorklund and R. D. Brown, "Physical play and cognitive development: Integrating activity, cognition, and education," *Child Development*, vol. 69, no. 3, pp. 604–606, 1998.

[378] B.-L. Zhong, S.-L. Chen, X. Tu, and Y. Conwell, "Loneliness and cognitive function in older adults: Findings from the chinese longitudinal healthy longevity survey," *The Journals of Gerontology: Series B*, vol. 72, no. 1, pp. 120–128, 2017.

[379] J. T. Cacioppo and S. Cacioppo, "The growing problem of loneliness," *The Lancet*, vol. 391, no. 10119, p. 426, Feb. 2018, publisher: Elsevier. [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)30142-9/abstract

[380] A. C. Barbour, "The impact of playground design on the play behaviors of children with differing levels of physical competence," *Early Childhood Research Quarterly*, vol. 14, no. 1, pp. 75 – 98, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885200699800076

[381] R. Bailey, K. Armour, D. Kirk, M. Jess, I. Pickup, R. Sandford, and BERA Physical Education and Sport Pedagogy Special Interest Group, "The

educational benefits claimed for physical education and school sport: an academic review," *Research Papers in Education*, vol. 24, no. 1, pp. 1–27, 2009. [Online]. Available: https://doi.org/10.1080/02671520701809817

[382] P. S. Strain and S. L. Odom, "Peer social initiations: Effective intervention for social skills development of exceptional children," *Exceptional Children*, vol. 52, no. 6, pp. 543–551, 1986.

[383] S. Bellini, *Building social relationships: A systematic approach to teaching social interaction skills to children and adolescents with autism spectrum disorders and other social difficulties.* AAPC Publishing, 2008.

[384] M. L. Combs and D. A. Slaby, "Social-skills training with children," in *Advances in clinical child psychology.* Springer, 1977, pp. 161–201.

[385] K. M. Lee, Y. Jung, J. Kim, and S. R. Kim, "Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction," *International Journal of Human-Computer Studies*, vol. 64, no. 10, pp. 962–973, 2006.

[386] F. Eyssel and N. Reich, "Loneliness makes the heart grow fonder (of robots)—on the effects of loneliness on psychological anthropomorphism," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 2013, pp. 121–122.

[387] N. Martelaro, V. C. Nneji, W. Ju, and P. Hinds, "Tell me more designing HRI to encourage more trust, disclosure, and companionship," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 2016, pp. 181–188.

[388] M. J. Matarić, "Socially assistive robotics: Human augmentation versus automation," *Science Robotics*, vol. 2, no. 4, p. eaam5410, 2017.

[389] Zoom Video Communications, Inc., "Zoom Video Conferencing Platform," https://zoom.us, 2023, accessed: 2025-07-01.

[390] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents," *International Journal of Social Robotics (IJSR)*, vol. 3, no. 1, pp. 41–52, Jan. 2011. [Online]. Available: https://doi.org/10.1007/s12369-010-0082-7

[391] A. Kristoffersson, S. Coradeschi, and A. Loutfi, "A review of mobile robotic telepresence," *Advances in Human-Computer Interaction*, 2013.

[392] J. M. Beer and L. Takayama, "Mobile remote presence systems for older adults: Acceptance, benefits, and concerns," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2011, pp. 19–26.

[393] A. M. Seelye, K. V. Wild, N. Larimer, S. Maxwell, P. Kearns, and J. A. Kaye, "Reactions to a remote-controlled video-communication robot in seniors' homes: A pilot study of feasibility and acceptance," *Telemedicine and e-Health*, vol. 18, no. 10, pp. 755–759, 2012, pMID: 23082794. [Online]. Available: https://doi.org/10.1089/tmj.2012.0026

[394] A. Cesta, G. Cortellessa, A. Orlandini, and L. Tiberio, "Long-term evaluation of a telepresence robot for the elderly: methodology and ecological case study," *International Journal of Social Robotics (IJSR)*, vol. 8, no. 3, pp. 421–441, 2016.

[395] F. Tanaka, T. Takahashi, S. Matsuzoe, N. Tazawa, and M. Morita, "Child-operated telepresence robot: A field trial connecting classrooms between australia and japan," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 5896–5901.

[396] ——, "Telepresence robot helps children in communicating with teachers who speak a different language," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2014, pp. 399–406.

[397] N. Soares, J. C. Kay, and G. Craven, "Mobile robotic telepresence solutions for the education of hospitalized children," *Perspectives in Health Information Management*, vol. 14, no. Fall, 2017.

[398] A. D. Pellegrini and P. K. Smith, "Physical activity play: The nature and function of a neglected aspect of play," *Child Development*, vol. 69, no. 3, pp. 577–598, 1998.

[399] X. Z. Tan, M. Vázquez, E. J. Carter, C. G. Morales, and A. Steinfeld, "Inducing bystander interventions during robot abuse with social mechanisms," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2018, pp. 169–177.

[400] R. Mahy, P. Matthews, and J. Rosenberg, "Traversal Using Relays around NAT (TURN): Relay Extensions to Session Traversal Utilities for NAT (STUN)," https://datatracker.ietf.org/doc/html/rfc5766, 2010.

[401] WebRTC Project, "WebRTC: Real-Time Communication for the Web," https://webrtc.org/, 2024, accessed: 2025-07-01.

[402] Google, "Protocol Buffers: Google's Data Interchange Format," https://developers.google.com/protocol-buffers, 2024, accessed: 2025-07-01.

[403] ——, "gRPC: A High Performance, Open-Source Universal RPC Framework," https://grpc.io/, 2024, accessed: 2025-07-01.

[404] Flutter. (2020) Flutter - beautiful native apps in record time. [Online]. Available: https://flutter.dev/

[405] Google, "Dart Programming Language," https://dart.dev/, 2024, accessed: 2025-07-01.

[406] J. C. Read, S. MacFarlane, and C. Casey, "Endurability, engagement and expectations: Measuring children's fun," in *Interaction Design and Children*, vol. 2. Shaker Publishing Eindhoven, 2002, pp. 1–23.

[407] H. Mendelson and K. Moon, "Modeling success and engagement for the app economy," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 569–578.

[408] D. Altig, S. Baker, J. M. Barrero, N. Bloom, P. Bunn, S. Chen, S. J. Davis, J. Leather, B. Meyer, E. Mihaylov *et al.*, "Economic uncertainty before and during the COVID-19 pandemic," *Journal of Public Economics*, p. 104274, 2020.

[409] B. Durodié, "Handling uncertainty and ambiguity in the COVID-19 pandemic." *Psychological Trauma: Theory, Research, Practice, and Policy*, vol. 12, no. S1, p. S61, 2020.

[410] L. Cabral and L. Xu, "Seller reputation and price gouging: Evidence from the COVID-19 pandemic," Mimeo, Tech. Rep., 2020.

[411] Center for Robot-Assisted Search and Rescue. (2020) Center for robot-assisted search and rescue (CRASAR). [Online]. Available: http://crasar.org/

[412] J. Driver IV, G. Davis, P. Ricciardelli, P. Kidd, E. Maxwell, and S. Baron-Cohen, "Gaze perception triggers reflexive visuospatial orienting," *Visual Cognition*, vol. 6, no. 5, pp. 509–540, 1999.

[413] S. Tiitinen and J. Ruusuvuori, "Using formulations and gaze to encourage parents to talk about their and their children's health and well-being," *Research on Language and Social Interaction*, vol. 47, no. 1, pp. 49–68, 2014.

[414] J. H. Wirth, D. F. Sacco, K. Hugenberg, and K. D. Williams, "Eye gaze as relational evaluation: Averted eye gaze leads to feelings of ostracism and relational devaluation," *Personality and Social Psychology Bulletin*, vol. 36, no. 7, pp. 869–882, 2010.

[415] R. Brooks and A. N. Meltzoff, "Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind," *Journal of Experimental Child Psychology*, vol. 130, pp. 67–78, 2015.

[416] E. A. von dem Hagen, R. S. Stoyanova, J. B. Rowe, S. Baron-Cohen, and A. J. Calder, "Direct gaze elicits atypical activation of the theory-of-mind network in autism spectrum conditions," *Cerebral Cortex*, vol. 24, no. 6, pp. 1485–1492, 2014.

[417] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, no. 2, pp. 70–76, 2006.

[418] D. C. Richardson and R. Dale, "Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension," *Cognitive Science*, vol. 29, no. 6, pp. 1045–1060, 2005.

[419] H. H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding," *Journal of Memory and Language*, vol. 50, no. 1, pp. 62–81, 2004.

[420] M. Tomasello and M. Carpenter, "Shared intentionality," *Developmental Science*, vol. 10, no. 1, pp. 121–125, 2007.

[421] T. Falck-Ytter, C. Carlström, and M. Johansson, "Eye contact modulates cognitive processing differently in children with autism," *Child Development*, vol. 86, no. 1, pp. 37–47, 2015.

[422] C. Chevallier, G. Kohls, V. Troiani, E. S. Brodkin, and R. T. Schultz, "The social motivation theory of autism," *Trends in Cognitive Sciences*, vol. 16, no. 4, pp. 231–239, 2012.

[423] T. Del Bianco, L. Mason, T. Charman, J. Tillman, E. Loth, H. Hayward, F. Shic, J. Buitelaar, M. H. Johnson, E. J. Jones *et al.*, "Temporal profiles of social attention are different across development in autistic and neurotypical people," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 6, no. 8, pp. 813–824, 2021.

[424] A. Senju and M. H. Johnson, "Atypical eye contact in autism: models, mechanisms and development," *Neuroscience & Biobehavioral Reviews*, vol. 33, no. 8, pp. 1204–1214, 2009.

[425] L. Krasny, B. J. Williams, S. Provencal, and S. Ozonoff, "Social skills interventions for the autism spectrum: Essential ingredients and a model curriculum," *Child and Adolescent Psychiatric Clinics*, vol. 12, no. 1, pp. 107–122, 2003.

[426] B. Hwang and C. Hughes, "The effects of social interactive training on early social communicative skills of children with autism," *Journal of Autism and Developmental Disorders (JADD)*, vol. 30, no. 4, pp. 331–343, 2000.

[427] J. T. Rapp, J. L. Cook, R. Nuta, C. Balagot, K. Crouchman, C. Jenkins, S. Karim, and C. Watters-Wybrow, "Further evaluation of a practitioner model for increasing eye contact in children with autism," *Behavior Modification*, vol. 43, no. 3, pp. 389–412, 2019.

[428] M. J. Weiss and T. Zane, "Three important things to consider when starting intervention for a child diagnosed with autism," *Behavior Analysis in Practice*, vol. 3, no. 2, p. 58, 2010.

[429] C. Whalen and L. Schreibman, "Joint attention training for children with autism using behavior modification procedures," *Journal of Child Psychology and Psychiatry*, vol. 44, no. 3, pp. 456–468, 2003.

[430] V. J. Carbone, L. O'Brien, E. J. Sweeney-Kerwin, and K. M. Albert, "Teaching eye contact to children with autism: A conceptual analysis and single case study," *Education and Treatment of Children*, vol. 36, no. 2, pp. 139–159, 2013.

[431] E. Anagnostou, N. Jones, M. Huerta, A. K. Halladay, P. Wang, L. Scahill, J. P. Horrigan, C. Kasari, C. Lord, D. Choi *et al.*, "Measuring social communication behaviors as a treatment endpoint in individuals with autism spectrum disorder," *Autism*, vol. 19, no. 5, pp. 622–636, 2015.

[432] H. McConachie, J. R. Parr, M. Glod, J. Hanratty, N. Livingstone, I. P. Oono, S. Robalino, G. Baird, B. Beresford, T. Charman *et al.*, "Systematic review of tools to measure outcomes for young children with autism spectrum disorder," *NIHR Health Technology Assessment Programme*, 2015.

[433] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system (ROS Melodic Morenia)." [Online]. Available: https://www.ros.org

[434] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018).* IEEE, 2018, pp. 59–66.

[435] H. Sloetjes and P. Wittenburg, "Annotation by category-ELAN and ISO DCR," in *6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008. [Online]. Available: https://archive.mpi.nl/tla/elan

[436] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4511–4520.

[437] A. C. MacPherson and C. Moore, "Attentional control by gaze cues in infancy," in *Gaze-Following: Its Development and Significance.* Psychology Press, 2017, pp. 53–75.

[438] T. W. Vernon, "Fostering a social child with autism: A moment-by-moment sequential analysis of an early social engagement intervention," *Journal of Autism and Developmental Disorders (JADD)*, vol. 44, pp. 3072–3082, 2014.

[439] P. Mundy and M. Crowson, "Joint attention and early social communication: Implications for research on intervention with autism," *Journal of Autism and Developmental Disorders (JADD)*, vol. 27, pp. 653–676, 1997.

[440] A. Masi, M. M. DeMayo, N. Glozier, and A. J. Guastella, "An overview of autism spectrum disorder, heterogeneity and treatment options," *Neuroscience bulletin*, vol. 33, no. 2, pp. 183–193, 2017.

[441] A. Speaks, "Autism statistics and facts." [Online]. Available: https://www.autismspeaks.org/autism-statistics

[442] R. D. Austin and G. P. Pisano, “Neurodiversity as a competitive advantage,” *Harvard Business Review*, vol. 95, no. 3, pp. 96–103, 2017.

[443] S. Lindsay, E. Cagliostro, M. Albarico, N. Mortaji, and L. Karon, “A systematic review of the benefits of hiring people with disabilities,” *Journal of Occupational Rehabilitation*, vol. 28, no. 4, pp. 634–655, 2018.

[444] M. Scott, A. Jacob, D. Hendrie, R. Parsons, S. Girdler, T. Falkmer, and M. Falkmer, “Employers’ perception of the costs and the benefits of hiring individuals with autism spectrum disorder in open employment in australia,” *PloS One*, vol. 12, no. 5, p. e0177607, 2017.

[445] C. D. Chisholm, E. K. Collison, D. R. Nelson, and W. H. Cordell, “Emergency department workplace interruptions are emergency physicians “interrupt-driven” and “multitasking”?” *Academic Emergency Medicine*, vol. 7, no. 11, pp. 1239–1243, 2000.

[446] L. Dabbish and R. E. Kraut, “Controlling interruptions: awareness displays and social motivation for coordination,” in *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 2004, pp. 182–191.

[447] M. Czerwinski, E. Horvitz, and S. Wilhite, “A diary study of task switching and interruptions,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2004, pp. 175–182.

[448] J. P. Borst, N. A. Taatgen, and H. van Rijn, “What makes interruptions disruptive? a process-model account of the effects of the problem state bottleneck on task interruption and resumption,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2971–2980.

[449] J. I. Westbrook, L. Li, T. D. Hooper, M. Z. Raban, S. Middleton, and E. C. Lehnbom, “Effectiveness of a ‘do not interrupt’bundled intervention to reduce interruptions during medication administration: a cluster randomised controlled feasibility study,” *BMJ quality & safety*, vol. 26, no. 9, pp. 734–742, 2017.

[450] L. Kenyon, “Managing autism in the workplace,” *Occupational Health & Well-being*, vol. 67, no. 6, p. 18, 2015.

[451] Easterseals Blog. (2020) What happens when my thought process is interrupted | Easterseals blog. [Online]. Available: https://blog.easterseals.com/aspergers-interrupted-japan-bullet-train/

[452] P. E. McKenna, I. Keller, J. L. Part, M. Y. Lim, R. Aylett, F. Broz, and G. Rajendran, “‘sorry to disturb you’ autism and robot interruptions,” in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 360–362.

[453] B. Scassellati, "How social robots will help us to diagnose, treat, and understand autism," in *Robotics Research*. Springer, 2007, pp. 552–563.

[454] M. J. Matarić and B. Scassellati, "Socially assistive robotics," *Springer Handbook of Robotics*, pp. 1973–1994, 2016.

[455] J. Wainer, K. Dautenhahn, B. Robins, and F. Amirabdollahian, "A pilot study with a novel setup for collaborative play of the humanoid robot KASPAR with children with autism," *International Journal of Social Robotics (IJSR)*, vol. 6, no. 1, pp. 45–65, 2014.

[456] P. T. Shattuck, S. C. Narendorf, B. Cooper, P. R. Sterzing, M. Wagner, and J. L. Taylor, "Postsecondary education and employment among youth with an autism spectrum disorder," *Pediatrics*, vol. 129, no. 6, pp. 1042–1049, 2012.

[457] R. L. Seaman and H. I. Cannella-Malone, "Vocational skills interventions for adults with autism spectrum disorder: A review of the literature," *Journal of Developmental and Physical Disabilities*, vol. 28, no. 3, pp. 479–494, 2016.

[458] M. J. Baker-Ericzén, A. C. Stahmer, and A. Burns, "Child demographics associated with outcomes in a community-based pivotal response training program," *Journal of positive behavior interventions*, vol. 9, no. 1, pp. 52–60, 2007.

[459] L. C. Eaves and H. H. Ho, "Young adult outcome of autism spectrum disorders," *Journal of Autism and Developmental Disorders (JADD)*, vol. 38, no. 4, pp. 739–747, 2008.

[460] T. Smith, L. Scahill, G. Dawson, D. Guthrie, C. Lord, S. Odom, S. Rogers, and A. Wagner, "Designing research studies on psychosocial interventions in autism," *Journal of Autism and Developmental Disorders (JADD)*, vol. 37, no. 2, pp. 354–366, 2007.

[461] J. B. Stoner, M. E. Angell, J. J. House, and S. J. Bock, "Transitions: Perspectives from parents of young children with autism spectrum disorder (ASD)," *Journal of Developmental and Physical Disabilities*, vol. 19, no. 1, pp. 23–39, 2007.

[462] L. Schreibman, "Intensive behavioral/psychoeducational treatments for autism: Research needs and future directions," *Journal of Autism and Developmental Disorders (JADD)*, vol. 30, no. 5, pp. 373–378, 2000.

[463] J. J. Donovan and D. J. Radosevich, "A meta-analytic review of the distribution of practice effect: Now you see it, now you don't." *Journal of Applied Psychology*, vol. 84, no. 5, p. 795, 1999.

[464] H. M. Hodgetts and D. M. Jones, "Interruption of the tower of london task: support for a goal-activation approach." *Journal of Experimental Psychology: General*, vol. 135, no. 1, p. 103, 2006.

[465] D. M. Cades, D. A. Boehm-Davis, J. G. Trafton, and C. A. Monk, "Mitigating disruptive effects of interruptions through training: What needs to be practiced?" *Journal of Experimental Psychology: Applied*, vol. 17, no. 2, p. 97, 2011.

[466] E. M. Altmann and J. G. Trafton, "Task interruption: Resumption lag and the role of cues," Michigan State University East Lansing Department of Psychology, Tech. Rep., 2004.

[467] D. D. Salvucci, N. A. Taatgen, and J. P. Borst, "Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 1819–1828.

[468] W. E. Jones and J. Moss, "Assessing the transfer of interruption resumption skill to novel tasks." *Journal of Experimental Psychology: Applied*, vol. 25, no. 2, p. 230, 2019.

[469] K. A. Latorella, "Investigating interruptions: An example from the flightdeck," in *Proceedings of the Human Factors and Ergonomics Society Annual meeting*, vol. 40, no. 4. SAGE Publications Sage CA: Los Angeles, CA, 1996, pp. 249–253.

[470] R. Dismukes, G. E. Young, R. L. Sumwalt III, and C. H. Null, "Cockpit interruptions and distractions: Effective management requires a careful balancing act," in *CRM Industry Conference*, 1998.

[471] M. Burke, R. Kraut, and D. Williams, "Social use of computer-mediated communication by adults on the autism spectrum," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 2010, pp. 425–434.

[472] P. T. Cummings, P. M. Fauchet, M. Goldfarb, M. W. Jones, M. Kunda, J. B. Perlin, N. Sarkar, K. G. Stassun, Z. E. Warren, and K. E. Zelik, "Engineering for inclusion: Empowering individuals with physical and neurological differences through engineering invention, research, and development," *Engineering*, 2020.

[473] M. J. Smith, E. J. Ginger, K. Wright, M. A. Wright, J. L. Taylor, L. B. Humm, D. E. Olsen, M. D. Bell, and M. F. Fleming, "Virtual reality job interview training in adults with autism spectrum disorder," *Journal of Autism and Developmental Disorders (JADD)*, vol. 44, no. 10, pp. 2450–2463, 2014.

[474] W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati, "The effect of presence on human-robot interaction," in *The 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2008, pp. 701–706.

[475] S. M. Srinivasan, I. K. Park, L. B. Neelly, and A. N. Bhat, "A comparison of the effects of rhythm and robotic interventions on repetitive behaviors and affective

states of children with autism spectrum disorder (ASD)," *Research in Autism Spectrum Disorders*, vol. 18, pp. 51–63, 2015.

[476] B. Robins, K. Dautenhahn, and P. Dickerson, "Embodiment and cognitive learning–can a humanoid robot help children with autism to learn about tactile social behaviour?" in *International Conference on Social Robotics*. Springer, 2012, pp. 66–75.

[477] S. Bruyère, H.-Y. Chang, and M. Saleh, "*Preliminary Report Summarizing the Results of Interviews and Focus Groups with Employers, Autistic Individuals, Service Providers, and Higher Education Career Counselors on Perceptions of Barriers and Facilitators for Neurodiverse Individuals in the Job Interview and Customer Interface Processes.*" 2020.

[478] S. M. Fiore, T. J. Wiltshire, E. J. Lobato, F. G. Jentsch, W. H. Huang, and B. Axelrod, "Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior," *Frontiers in Psychology*, vol. 4, p. 859, 2013.

[479] Microsoft. (2020) Azure Kinect DK – develop AI models | Microsoft Azure. [Online]. Available: https://azure.microsoft.com/en-us/services/kinect-dk/

[480] A. Farhadi and J. Redmon, "YOLOv3: An incremental improvement," in *Computer Vision and Pattern Recognition*, 2018, pp. 1804–02 767.

[481] TeamViewer, "Teamviewer remote desktop software." [Online]. Available: https://www.teamviewer.com/

[482] AnyDesk, "Anydesk remote desktop software." [Online]. Available: https://anydesk.com/

[483] P. Lally, C. H. Van Jaarsveld, H. W. Potts, and J. Wardle, "How are habits formed: Modelling habit formation in the real world," *European journal of social psychology*, vol. 40, no. 6, pp. 998–1009, 2010.

[484] Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, "ELAN." [Online]. Available: https://archive.mpi.nl/tla/elan

[485] R. Vollmeyer and F. Rheinberg, "Does motivation affect performance via persistence?" *Learning and instruction*, vol. 10, no. 4, pp. 293–309, 2000.

[486] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation models defining a new era in vision: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[487] Capgemini, "Why consumers love generative AI: report from the Capgemini Research Institute," 2023, accessed: 2024-08-08. [Online]. Available: https://www.capgemini.com/insights/research-library/creative-and-generative-ai/

[488] A. d. Garcez and L. C. Lamb, "Neurosymbolic AI: the 3rd wave," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12 387–12 406, 2023.

[489] K. Wang, Y. Lu, M. Santacroce, Y. Gong, C. Zhang, and Y. Shen, "Adapting LLM agents through communication," *arXiv preprint arXiv:2310.01444*, 2023.

[490] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang *et al.*, "When large language models meet personalization: Perspectives of challenges and opportunities," *World Wide Web*, vol. 27, no. 4, p. 42, 2024.

[491] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, and S. Tellex, "Planning with large language models via corrective re-prompting," in *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[492] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[493] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[494] O. Biza, D. Wang, R. Platt, J.-W. van de Meent, and L. L. Wong, "Action priors for large action spaces in robotics," *arXiv preprint arXiv:2101.04178*, 2021.

[495] L. Giray, "Prompt engineering with ChatGPT: a guide for academic writers," *Annals of Biomedical Engineering*, vol. 51, no. 12, pp. 2629–2633, 2023.

[496] B. Meskó, "Prompt engineering as an important emerging skill for medical professionals: tutorial," *Journal of Medical Internet Research*, vol. 25, p. e50638, 2023.

[497] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with ChatGPT," *arXiv preprint arXiv:2302.11382*, 2023.

[498] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[499] Q. Huang, X. Liu, T. Ko, B. Wu, W. Wang, Y. Zhang, and L. Tang, "Selective prompting tuning for personalized conversations with LLMs," *arXiv preprint arXiv:2406.18187*, 2024.

[500] P. Denny, V. Kumar, and N. Giacaman, "Conversing with copilot: Exploring prompt engineering for solving CS1 problems using natural language," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V.1*, 2023, pp. 1136–1142.

[501] Y. Dong, R. Mu, G. Jin, Y. Qi, J. Hu, X. Zhao, J. Meng, W. Ruan, and X. Huang, “Building guardrails for large language models,” *arXiv preprint arXiv:2402.01822*, 2024.

[502] M. Niknazar, P. V. Haley, L. Ramanan, S. T. Truong, Y. Shrinivasan, A. K. Bhowmick, P. Dey, A. Jagmohan, H. Maheshwari, S. Ponoth *et al.*, “Building a domain-specific guardrail model in production,” *arXiv preprint arXiv:2408.01452*, 2024.

[503] I. Leite, C. Martinho, and A. Paiva, “Social robots for long-term interaction: a survey,” *International Journal of Social Robotics (IJSR)*, vol. 5, pp. 291–308, 2013.

[504] K. Lyu, H. Zhao, X. Gu, D. Yu, A. Goyal, and S. Arora, “Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates,” *arXiv preprint arXiv:2402.18540*, 2024.

[505] J.-T. Chen and C.-M. Huang, “Forgetful large language models: Lessons learned from using LLMs in robot programming,” in *Proceedings of the AAAI Symposium Series*, vol. 2, no. 1, 2023, pp. 508–513.

[506] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, “A review of safe reinforcement learning: Methods, theory and applications,” *arXiv preprint arXiv:2205.10330*, 2022.

[507] J. Garcıa and F. Fernández, “A comprehensive survey on safe reinforcement learning,” *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[508] J. Brawer, D. Ghose, K. Candon, M. Qin, A. Roncone, M. Vázquez, and B. Scassellati, “Interactive policy shaping for human-robot collaboration with transparent matrix overlays,” in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2023, pp. 525–533.

[509] S. K. Paul, “Continually learning planning agent for large environments guided by LLMs,” in *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 377–382.

[510] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, “A comprehensive overview of large language models,” *arXiv preprint arXiv:2307.06435*, 2023.

[511] S. M. Shieber, *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. MIT Press, 2004.

[512] S.-C. Chen, C. Jones, and W. Moyle, “Social robots for depression in older adults: a systematic review,” *Journal of Nursing Scholarship*, vol. 50, no. 6, pp. 612–622, 2018.

[513] J. Coupland, *Small talk*. Routledge, 2014.

[514] J. Laver, "Linguistic routines and politeness in greeting and parting," *Conversational Routine*, vol. 289304, 1981.

[515] S. Eggins and D. Slade, *Analysing Casual Conversation*. Equinox Publishing Ltd., 2004.

[516] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[517] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[518] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[519] R. Ramnauth, D. Brščić, and B. Scassellati, "A robot-assisted approach to small-talk training for adults with ASD," Jun 2024. [Online]. Available: osf.io/kad27

[520] K. Tracy and N. Coupland, *Multiple goals in discourse*. Multilingual Matters, 1990.

[521] M. Yeomans, M. E. Schweitzer, and A. W. Brooks, "The conversational circumplex: Identifying, prioritizing, and pursuing informational and relational motives in conversation," *Current Opinion in Psychology*, vol. 44, pp. 293–302, 2022.

[522] J. Coupland, "Introduction: Sociolinguistic perspectives on small talk," in *Small Talk*. Routledge, 2014, pp. 1–25.

[523] R. Ramnauth, "Small talk for social robots," Feb 2024. [Online]. Available: osf.io/6d9z2

[524] OpenAI, "OpenAI ChatGPT API," 2024. [Online]. Available: https://platform.openai.com/

[525] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014, pp. 216–225.

[526] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[527] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[528] S. Palan and C. Schitter, "Prolific. a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[529] C.-C. Lin, A. Y. Huang, and O. H. Lu, "Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review," *Smart Learning Environments*, vol. 10, no. 1, p. 41, 2023.

[530] K. Fischer, K. S. Lohan, C. Nehaniv, and H. Lehmann, "Effects of different kinds of robot feedback," in *Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5*. Springer, 2013, pp. 260–269.

[531] M. Price, K. Handley, J. Millar, and B. O'donovan, "Feedback: all that effort, but what is the effect?" *Assessment & Evaluation in Higher Education*, vol. 35, no. 3, pp. 277–289, 2010.

[532] Z.-W. Hong, I. Shenfeld, T.-H. Wang, Y.-S. Chuang, A. Pareja, J. Glass, A. Srivastava, and P. Agrawal, "Curiosity-driven red-teaming for large language models," *arXiv preprint arXiv:2402.19464*, 2024.

[533] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[534] K. Hurlbutt and L. Chalmers, "Employment and adults with asperger syndrome," *Focus on autism and other developmental disabilities*, vol. 19, no. 4, pp. 215–222, 2004.

[535] A. M. Roux, P. T. Shattuck, B. P. Cooper, K. A. Anderson, M. Wagner, and S. C. Narendorf, "Postsecondary employment experiences among young adults with an autism spectrum disorder," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 52, no. 9, pp. 931–939, 2013.

[536] J. Holmes, "When small talk is a big deal: Sociolinguistic challenges in the workplace," *Second Language Needs Analysis*, vol. 344, p. 371, 2005.

[537] B. Joseph, K. B. Kearney, M. P. Brady, A. Downey, and A. Torres, "Teaching small talk: Increasing on-topic conversational exchanges in college students with intellectual and developmental disabilities using remote audio coaching," *Behavior Modification*, vol. 45, no. 2, pp. 251–271, 2021.

[538] A. Ohl, M. Grice Sheff, S. Small, J. Nguyen, K. Paskor, and A. Zanjirian, "Predictors of employment status among adults with autism spectrum disorder," *Work*, vol. 56, no. 2, pp. 345–355, 2017.

[539] P. Pullin, "Small talk, rapport, and international communicative competence: Lessons to learn from BELF," *The Journal of Business Communication (1973)*, vol. 47, no. 4, pp. 455–476, 2010.

[540] S. Bristoll and J. Dickinson, "Small talk, big results," *Newsli*, vol. 92, pp. 6–13, 2015.

[541] V. Garrels, "Getting good at small talk: Student-directed learning of social conversation skills," *European Journal of Special Needs Education*, vol. 34, no. 3, pp. 393–402, 2019.

[542] J. Nadel, O. Grynszpan, and J.-C. Martin, "Autism and socially interactive agents," in *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, 2022, pp. 437–462.

[543] H. Kozima, C. Nakagawa, and Y. Yasuda, "Interactive robots for communication-care: A case-study in autism therapy," in *IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2005, pp. 341–346.

[544] B. Robins, K. Dautenhahn, and P. Dickerson, "From isolation to communication: a case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot," in *2009 Second International Conferences on Advances in Computer-Human Interactions*. IEEE, 2009, pp. 205–211.

[545] J. Kanero, V. Geçkin, C. Oranç, E. Mamus, A. C. Küntay, and T. Göksun, "Social robots for early language learning: Current evidence and future directions," *Child Development Perspectives*, vol. 12, no. 3, pp. 146–151, 2018.

[546] F. Babel, J. Kraus, L. Miller, M. Kraus, N. Wagner, W. Minker, and M. Baumann, "Small talk with a robot? the impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity," *International Journal of Social Robotics (IJSR)*, pp. 1–14, 2021.

[547] N. Ise and T. Iio, "Social robot encouraging two strangers to talk with each other for their relationships," in *Companion of the 2021 ACM/IEEE International Conference on HRI*, 2021, pp. 144–147.

[548] E. Nichols, S. R. Siskind, L. Ivanchuk, G. Pérez, W. Kamino, S. Šabanović, and R. Gomez, "Hey haru, let's be friends! using the tiers of friendship to build rapport through small talk with the tabletop robot haru," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 6101–6108.

[549] P. Pexman, L. Reggin, and K. Lee, "Addressing the challenge of verbal irony: Getting serious about sarcasm training," *Languages*, vol. 4, no. 2, p. 23, 2019.

[550] K. Silver and S. Parsons, "Perspectives of autistic adults on the strategies that help or hinder successful conversations," *Autism & Developmental Language Impairments*, vol. 7, p. 23969415221101113, 2022.

[551] S. Bellini and J. K. Peters, "Social skills training for youth with autism spectrum disorders," *Child and Adolescent Psychiatric Clinics of North America*, vol. 17, no. 4, pp. 857–873, 2008.

[552] J. S. Akers, N. Pyle, T. S. Higbee, D. Pyle, and K. R. Gerencser, "A synthesis of script fading effects with individuals with autism spectrum disorder: A 20-year review," *Review Journal of Autism and Developmental Disorders*, vol. 3, pp. 1–17, 2016.

[553] F. J. DeMatteo, P. S. Arter, C. Sworen-Parise, M. Fasciana, and M. A. Paulhamus, "Social skills training for young adults with autism spectrum disorder: Overview and implications for practice." *National Teacher Education Journal*, vol. 5, no. 4, 2012.

[554] S. Bejerot, J. M. Eriksson, and E. Mörtberg, "Social anxiety in adult autism spectrum disorder," *Psychiatry research*, vol. 220, no. 1-2, pp. 705–707, 2014.

[555] E. A. Perkins and K. A. Berkman, "Into the unknown: Aging with autism spectrum disorders," *American Journal on Intellectual and Developmental Disabilities*, vol. 117, no. 6, pp. 478–496, 2012.

[556] P. F. Gerhardt and I. Lainer, "Addressing the needs of adolescents and adults with autism: A crisis on the horizon," *Journal of Contemporary Psychotherapy*, vol. 41, no. 1, pp. 37–45, 2011.

[557] A. Jaworski, "Silence and small talk," in *Small talk*. Routledge, 2014, pp. 110–132.

[558] C. Duggal, B. Dua, R. Chokhani, and K. Sengupta, "What works and how: Adult learner perspectives on an autism intervention training program in India," *Autism*, vol. 24, no. 1, pp. 246–257, 2020.

[559] L. Wermer, M. E. Brock, and R. L. Seaman, "Efficacy of a teacher training a paraprofessional to promote communication for a student with autism and complex communication needs," *Focus on Autism and Other Developmental Disabilities*, vol. 33, no. 4, pp. 217–226, 2018.

[560] M. Dehqonova and G. Tagonova, "The importance of didactic games in speech therapy in the development of speech in children with autism and the ability to choose effective methods," *European Journal of Innovation In Nonformal Education*, vol. 2, no. 1, pp. 133–137, 2022.

[561] D. J. Delprato, "Comparisons of discrete-trial and normalized behavioral language intervention for young children with autism," *Journal of Autism and Developmental Disorders (JADD)*, vol. 31, pp. 315–325, 2001.

[562] R. J. Christensen-Sandfort and S. B. Whinnery, “Impact of milieu teaching on communication skills of young children with autism spectrum disorder,” *Topics in Early Childhood Special Education*, vol. 32, no. 4, pp. 211–222, 2013.

[563] G. R. Mancil, “Milieu therapy as a communication intervention: A review of the literature related to children with autism spectrum disorder,” *Education and Training in Developmental Disabilities*, pp. 105–117, 2009.

[564] A. P. Kaiser, T. B. Hancock, and J. P. Nietfeld, “The effects of parent-implemented enhanced milieu teaching on the social communication of children who have autism,” *Early Education and Development*, vol. 11, no. 4, pp. 423–446, 2000.

[565] B. T. Ogletree, J. R. Price, and J. M. Campbell, “A selective review of milieu interventions for adults with severe intellectual disabilities,” *Current Developmental Disorders Reports*, vol. 7, pp. 109–115, 2020.

[566] D. Leyzberg, E. Avrunin, J. Liu, and B. Scassellati, “Robots that express emotion elicit better human teaching,” in *Proceedings of the 6th International Conference on Human-Robot Interaction (HRI)*, 2011, pp. 347–354.

[567] R. Loomes, L. Hull, and W. P. L. Mandy, “What is the male-to-female ratio in autism spectrum disorder? a systematic review and meta-analysis,” *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 56, no. 6, pp. 466–474, 2017.

[568] R. Paul, S. M. Orlovski, H. C. Marcinko, and F. Volkmar, “Conversational behaviors in youth with high-functioning ASD and asperger syndrome,” *Journal of Autism and Developmental Disorders (JADD)*, vol. 39, pp. 115–125, 2009.

[569] V. Petrolini, M. Jorba, and A. Vicente, “What does it take to be rigid? reflections on the notion of rigidity in autism,” *Frontiers in Psychiatry*, vol. 14, p. 1072362, 2023.

[570] A. Nadig, I. Lee, L. Singh, K. Bosshart, and S. Ozonoff, “How does the topic of conversation affect verbal exchange and eye gaze? a comparison between typical development and high-functioning autism,” *Neuropsychologia*, vol. 48, no. 9, pp. 2730–2739, 2010.

[571] R. A. Schwartz-Mette and A. J. Rose, “Conversational self-focus in adolescent friendships: Observational assessment of an interpersonal process and relations with internalizing symptoms and friendship quality,” *Journal of Social and Clinical Psychology*, vol. 28, no. 10, pp. 1263–1297, 2009.

[572] K. Ochi, N. Ono, K. Owada, M. Kojima, M. Kuroda, S. Sagayama, and H. Yamasue, “Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder,” *PloS One*, vol. 14, no. 12, p. e0225377, 2019.

[573] T. B. van Ommeren, S. Begeer, A. M. Scheeren, and H. M. Koot, "Measuring reciprocity in high functioning children and adolescents with autism spectrum disorders," *Journal of Autism and Developmental Disorders (JADD)*, vol. 42, pp. 1001–1010, 2012.

[574] L. K. Koegel, M. N. Park, and R. L. Koegel, "Using self-management to improve the reciprocal social conversation of children with autism spectrum disorder," *Journal of Autism and Developmental Disorders (JADD)*, vol. 44, pp. 1055–1063, 2014.

[575] C. Ying Sng, M. Carter, and J. Stephenson, "A systematic review of the comparative pragmatic differences in conversational skills of individuals with autism," *Autism & Developmental Language Impairments*, vol. 3, p. 2396941518803806, 2018.

[576] H. M. Geurts, B. Corbett, and M. Solomon, "The paradox of cognitive flexibility in autism," *Trends in Cognitive Sciences*, vol. 13, no. 2, pp. 74–82, 2009.

[577] L. Balderaz, "Social skills interventions for adults with ASD: a review of the literature," *Journal of Psychosocial Rehabilitation and Mental Health*, vol. 7, no. 1, pp. 45–54, 2020.

[578] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[579] W. Schneider and A. D. Fisk, "Attention theory and mechanisms for skilled performance," in *Advances in Psychology.* Elsevier, 1983, vol. 12, pp. 119–143.

[580] B. J. Zimmerman, "Development and adaptation of expertise: The role of self-regulatory processes and beliefs," *The Cambridge Handbook of Expertise and Expert Performance*, vol. 186, pp. 705–722, 2006.

[581] S. H. Spence, "Social skills training with children and young people: Theory, evidence and practice," *Child and adolescent mental health*, vol. 8, no. 2, pp. 84–96, 2003.

[582] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[583] V. Lim, M. Rooksby, and E. S. Cross, "Social robots on a global stage: establishing a role for culture during human–robot interaction," *International Journal of Social Robotics (IJSR)*, vol. 13, no. 6, pp. 1307–1333, 2021.

[584] K. Scaler Scott, J. A. Tetnowski, J. R. Flaitz, and J. S. Yaruss, "Preliminary study of disfluency in school-aged children with autism," *International Journal of Language & Communication Disorders*, vol. 49, no. 1, pp. 75–89, 2014.

[585] E. Watson, T. Viana, and S. Zhang, "Machine learning driven developments in behavioral annotation: A recent historical review," *International Journal of Social Robotics (IJSR)*, pp. 1–14, 2024.

[586] Centers for Disease Control and Prevention, "Prevalence of parent-reported diagnoses of ASD among children," 2018. [Online]. Available: https://www.cdc.gov/nchs/products/databriefs/db291.htm

[587] ——, "Developmental disabilities," 2023. [Online]. Available: https://www.cdc.gov/environmental-health-tracking/php/data-research/developmental-disabilities.html

[588] N. Doroud, M. Cappy, K. Grant *et al.*, "Sensory rooms within mental health settings," *Occupational Therapy in Mental Health*, pp. 1–21, 2024.

[589] M. Biondi, M. Pasquini, L. Tarsitani *et al.*, *Empathy, Normalization and De-escalation*. Springer, 2021.

[590] T. E. Scruggs, M. A. Mastropieri, and K. A. McDuffie, "Co-teaching in inclusive classrooms: A metasynthesis of qualitative research," *Exceptional Children*, vol. 73, no. 4, pp. 392–416, 2007.

[591] J. Grace, "Multisensory rooms: characteristics and barriers to effective practice," *Tizard Learn. Disabil. Rev.*, vol. 25, no. 2, pp. 67–75, 2020.

[592] National Council for Special Education, "Sensory spaces in schools," 2021. [Online]. Available: https://ncse.ie/wp-content/uploads/2021/10/NCSE-Sensory-Spaces-in-Schools-2021.pdf

[593] W. Johal, "Research trends in social robots for learning," *Current Robotics Reports*, vol. 1, no. 3, pp. 75–83, 2020.

[594] S. Rasouli, G. Gupta, E. Nilsen, and K. Dautenhahn, "Potential applications of social robots in robot-assisted interventions for social anxiety," *International Journal of Social Robotics (IJSR)*, vol. 14, no. 5, pp. 1–32, 2022.

[595] United Federation of Teachers, "UFT helps provide sensory tool kits to more than 16,000 NYC classrooms," 2022. [Online]. Available: https://www.uft.org/news/press-releases/uft-helps-provide-sensory-tool-kits-more-16000-nyc-classrooms

[596] R. Ros and P. A. Graziano, "A transdiagnostic examination of self-regulation: Comparisons across preschoolers with ASD, ADHD, and typically developing children," *Journal of Clinical Child Adolescent Psychology*, vol. 49, no. 4, pp. 493–508, 2020.

[597] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learning and instruction*, vol. 4, no. 4, pp. 295–312, 1994.

[598] K. Fraser, D. MacKenzie, and J. Versnel, "Complex trauma in children and youth," *Occupational Therapy in Mental Health*, vol. 33, pp. 199–216, 2017.

[599] F. Dino, R. Zandie, H. Abdollahi, S. Schoeder, and M. H. Mahoor, "Delivering cognitive behavioral therapy using a conversational social robot," in *IROS*. IEEE, 2019, pp. 2089–2095.

[600] S. Papert, *Children, Computers & Powerful Ideas.* Harvester, 1980.

[601] J. Piaget, *The Psychology of Intelligence.* Routledge, 2005.

[602] E. L. Deci and R. M. Ryan, "Self-determination theory," *Handbook of Theories of Social Psychology*, vol. 1, no. 20, pp. 416–436, 2012.

[603] B. J. Zimmerman, "Becoming a self-regulated learner: An overview," *Theory into Practice*, vol. 41, no. 2, pp. 64–70, 2002.

[604] P. K. Smith, "A longitudinal study of social participation in preschool children," *Developmental Psychology*, vol. 14, no. 5, p. 517, 1978.

[605] J. A. Pardede, S. F. A. Sitepu, and M. Saragih, "The influence of deep breath relaxation techniques and five-finger hypnotic therapy on preoperative patient anxiety," *Journal of Psychiatry*, vol. 3, pp. 1–8, 2018.

[606] United States Department of Education, "Individuals with disabilities education act," 2004. [Online]. Available: https://sites.ed.gov/idea/

[607] E. H. Schein, *Helping: How to Offer, Give, and Receive Help.* San Francisco: Berrett-Koehler Pub., 2009.

[608] H. Beilin, *The Development of Prosocial Behavior.* Academic Press, 2013.

[609] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, K. L. Koay, and I. Werry, "What is a robot companion—friend, assistant or butler?" in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 1192–1197.

[610] N. Ezer, A. D. Fisk, and W. A. Rogers, "Attitudinal and intentional acceptance of domestic robots by younger and older adults," in *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments: 5th International Conference, UAHCI 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009. Proceedings, Part II 5.* Springer, 2009, pp. 39–48.

[611] A. C. Horstmann and N. C. Krämer, "Great expectations? relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment," *Frontiers in Psychology*, vol. 10, p. 939, 2019.

[612] E. Ganal, M. Habenicht, and B. Lugrin, "Excuse me, may I disturb you? the influence of politeness of a social robot on the perception of interruptions," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2024, pp. 2240–2247.

[613] E. Lumer and H. Buschmeier, "Should robots be polite? expectations about politeness in human–robot interaction," *Frontiers in Robotics and AI*, vol. 10, p. 1242127, 2023.

[614] S. Akheel, "Guardrails for large language models: A review of techniques and challenges," *Journal of Artificial Intelligence, Machine Learning and Data Science*, vol. 3, no. 1, pp. 2504–2512, 2025.

[615] R. Pakkar, C. Clabaugh, R. Lee, E. Deng, and M. J. Matarićć, "Designing a socially assistive robot for long-term in-home use for children with autism spectrum disorders," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[616] N. Ackovska and B. Ilijoski, "Online availability of a robot therapy for children with autism," in *SoutheastCon 2023*. IEEE, 2023, pp. 603–609.

[617] D. A. Urdanivia Alarcon, S. Cano, F. H. R. Paucar, R. F. P. Quispe, F. Talavera-Mendoza, and M. E. R. Zegarra, "Exploring the effect of robot-based video interventions for children with autism spectrum disorder as an alternative to remote education," *Electronics*, vol. 10, no. 21, p. 2577, 2021.

[618] J. Albo-Canals, M. Heerink, M. Diaz, V. Padillo, M. Maristany, A. Barco, C. Angulo, A. Riccio, L. Brodsky, S. Dufresne *et al.*, "Comparing two LEGO robotics-based interventions for social skills training with children with ASD," in *The 22nd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2013, pp. 638–643.

[619] S. Ali, F. Mehmood, D. Dancey, Y. Ayaz, M. J. Khan, N. Naseer, R. D. C. Amadeu, H. Sadia, and R. Nawaz, "An adaptive multi-robot therapy for improving joint attention and imitation of ASD children," *IEEE Access*, vol. 7, pp. 81 808–81 825, 2019.

[620] S. Ali, F. Mehmood, M. J. Khan, Y. Ayaz, U. Asgher, H. Sadia, E. Edifor, and R. Nawaz, "A preliminary study on effectiveness of a standardized multi-robot therapy for improvement in collaborative multi-human interaction of children with ASD," *IEEE Access*, vol. 8, pp. 109 466–109 474, 2020.

[621] S. Ali, F. Mehmood, Y. Ayaz, M. J. Khan, H. Sadia, and R. Nawaz, "Comparing the effectiveness of different reinforcement stimuli in a robotic therapy for children with ASD," *IEEE Access*, vol. 8, pp. 13 128–13 137, 2020.

[622] S. Ali, F. Mehmood, Y. Ayaz, M. Sajid, H. Sadia, and R. Nawaz, "An experimental trial: multi-robot therapy for categorization of autism level using hidden

markov model," *Journal of Educational Computing Research*, vol. 60, no. 3, pp. 722–741, 2022.

[623] A. Al-Nafjan, N. Alhakbani, and A. Alabdulkareem, "Measuring engagement in robot-assisted therapy for autistic children," *Behavioral Sciences*, vol. 13, no. 8, p. 618, 2023.

[624] F. Alnajjar, M. Cappuccio, A. Renawi, O. Mubin, and C. K. Loo, "Personalized robot interventions for autistic children: an automated methodology for attention assessment," *International Journal of Social Robotics (IJSR)*, vol. 13, no. 1, pp. 67–82, 2021.

[625] F. Amirabdollahian, B. Robins, K. Dautenhahn, and Z. Ji, "Investigating tactile event recognition in child-robot interaction for use in autism therapy," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 5347–5351.

[626] A. Amirova, N. Rakhymbayeva, A. Zhanatkyzy, Z. Telisheva, and A. Sandygulova, "Effects of parental involvement in robot-assisted autism therapy," *Journal of Autism and Developmental Disorders (JADD)*, vol. 53, no. 1, pp. 438–455, 2023.

[627] C. A. Pop, R. Simut, S. Pintea, J. Saldien, A. Rusu, D. David, J. Vanderfaeillie, D. Lefeber, and B. Vanderborght, "Can the social robot Probo help children with autism to identify situation-based emotions? a series of single case experiments," *International Journal of Humanoid Robotics*, vol. 10, no. 03, p. 1350025, 2013.

[628] H. E. Andreae, P. M. Andreae, J. Low, and D. Brown, "A study of Auti: a socially assistive robotic toy," in *Proceedings of the 2014 Conference on Interaction Design and Children*, 2014, pp. 245–248.

[629] S. Annunziata, L. Santos, A. Caglio, A. Geminiani, E. Brazzoli, E. Piazza, I. Olivieri, A. Pedrocchi, and A. Cavallini, "Interactive mirrOring Games wIth sOCial rObot (IOGIOCO): a pilot study on the use of intransitive gestures in a sample of italian preschool children with autism spectrum disorder," *Frontiers in Psychiatry*, vol. 15, p. 1356331, 2024.

[630] S. M. Anzalone, E. Tilmont, S. Boucenna, J. Xavier, A.-L. Jouen, N. Bodeau, K. Maharatna, M. Chetouani, D. Cohen, and MICHELANGELO Study Group and others, "How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D+ time) environment during a joint attention induction task with a robot," *Research in Autism Spectrum Disorders*, vol. 8, no. 7, pp. 814–826, 2014.

[631] S. Attawibulkul, N. Asawalertsak, P. Suwawong, P. Wattanapongsakul, W. Jutharee, and B. Kaewkamnerdpong, "Using a daily routine game on the bliss robot for supporting personal-social development in children with autism

and other special needs," in *2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. IEEE, 2019, pp. 695–700.

[632] M. Axelsson, M. Racca, D. Weir, and V. Kyrki, "A participatory design process of a robotic tutor of assistive sign language for children with autism," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–8.

[633] A. A. Aziz, F. F. M. Moganan, A. Ismail, and A. M. Lokman, "Autistic children's kansei responses towards humanoid-robot as teaching mediator," *Procedia Computer Science*, vol. 76, pp. 488–493, 2015.

[634] K. Baraka, F. S. Melo, M. Couto, and M. Veloso, "Optimal action sequence generation for assistive agents in fixed horizon tasks," *Autonomous Agents and Multi-Agent Systems*, vol. 34, no. 2, p. 33, 2020.

[635] K. Baraka, M. Couto, F. S. Melo, A. Paiva, and M. Veloso, "Sequencing matters: Investigating suitable action sequences in robot-assisted autism therapy," *Frontiers in Robotics and AI*, vol. 9, p. 784249, 2022.

[636] E. I. Barakova, P. Bajracharya, M. Willemsen, T. Lourens, and B. Huskens, "Long-term lego therapy with humanoid robot for children with ASD," *Expert Systems*, vol. 32, no. 6, pp. 698–709, 2015.

[637] J. A. Barnes, C. H. Park, A. Howard, and M. Jeon, "Child-robot interaction in a musical dance game: An exploratory comparison study between typically developing children and children with autism," *International Journal of Human–Computer Interaction*, vol. 37, no. 3, pp. 249–266, 2021.

[638] M. Begum, R. W. Serna, D. Kontak, J. Allspaw, J. Kuczynski, H. A. Yanco, and J. Suarez, "Measuring the efficacy of robots in autism therapy: How informative are standard HRI metrics?" in *Proceedings of the tenth annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015, pp. 335–342.

[639] T. Esubalew, U. Lahiri, A. R. Swanson, J. A. Crittendon, Z. E. Warren, N. Sarkar *et al.*, "A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 2, pp. 289–299, 2012.

[640] E. Bekele, U. Lahiri, J. Davidson, Z. Warren, and N. Sarkar, "Development of a novel robot-mediated adaptive response system for joint attention task for children with autism," in *20th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2011, pp. 276–281.

[641] E. Bekele, J. A. Crittendon, A. Swanson, N. Sarkar, and Z. E. Warren, "Pilot clinical application of an adaptive robotic system for young children with autism," *Autism*, vol. 18, no. 5, pp. 598–608, 2014.

[642] I. van den Berk-Smeekens, M. van Dongen-Boomsma, M. W. De Korte, J. C. Den Boer, I. J. Oosterling, N. C. Peters-Scheffer, J. K. Buitelaar, E. I. Barakova, T. Lourens, W. G. Staal *et al.*, "Adherence and acceptability of a robot-assisted pivotal response treatment protocol for children with autism spectrum disorder," *Scientific Reports*, vol. 10, no. 1, p. 8110, 2020.

[643] J. Bharatharaj, L. Huang, A. M. Al-Jumaily, C. Krageloh, and M. R. Elara, "Effects of adapted model-rival method and parrot-inspired robot in improving learning and social interaction among children with autism," in *2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*. IEEE, 2016, pp. 1–5.

[644] J. Bharatharaj, L. Huang, A. Al-Jumaily, M. R. Elara, and C. Krägeloh, "Investigating the effects of robot-assisted therapy among children with autism spectrum disorder using bio-markers," in *IOP Conference Series: Materials Science and Engineering*, vol. 234, no. 1. IOP Publishing, 2017, p. 012017.

[645] J. Bharatharaj, L. Huang, R. E. Mohan, A. Al-Jumaily, and C. Krägeloh, "Robot-assisted therapy for learning and social interaction of children with autism spectrum disorder," *Robotics*, vol. 6, no. 1, p. 4, 2017.

[646] J. Bharatharaj, L. Huang, A. Al-Jumaily, R. E. Mohan, and C. Krägeloh, "Sociopsychological and physiological effects of a robot-assisted therapy for children with autism," *International Journal of Advanced Robotic Systems*, vol. 14, no. 5, p. 1729881417736895, 2017.

[647] G. Bird, J. Leighton, C. Press, and C. Heyes, "Intact automatic imitation of human and robot actions in autism spectrum disorders," *Proceedings of the Royal Society B: Biological Sciences*, vol. 274, no. 1628, pp. 3027–3031, 2007.

[648] L. Boccanfuso, S. Scarborough, R. K. Abramson, A. V. Hall, H. H. Wright, and J. M. O'Kane, "A low-cost socially assistive robot and robot-assisted intervention for children with autism spectrum disorder: field trials and lessons learned," *Autonomous Robots*, vol. 41, no. 3, pp. 637–655, 2017.

[649] M. Brienza, F. Laus, V. Guglielmi, G. Carriero, M. Sileo, M. Grisolia, G. Palermo, D. D. Bloisi, F. Pierri, M. Turi *et al.*, "HRI-based gaze-contingent eye tracking for autism spectrum disorder treatment: A preliminary study using a NAO robot," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 2665–2670.

[650] N. Bugnariu, C. Young, K. Rockenbach, R. M. Patterson, C. Garver, I. Ranatunga, M. Beltran, N. Torres-Arenas, and D. Popa, "Human-robot interaction as a tool to evaluate and quantify motor imitation behavior in children with autism spectrum disorders," in *2013 International Conference on Virtual Rehabilitation (ICVR)*. IEEE, 2013, pp. 57–62.

[651] H. Cai, Y. Fang, Z. Ju, C. Costescu, D. David, E. Billing, T. Ziemke, S. Thill, T. Belpaeme, B. Vanderborght *et al.*, "Sensing-enhanced therapy system for assessing children with autism spectrum disorders: A feasibility study," *IEEE Sensors Journal*, vol. 19, no. 4, pp. 1508–1518, 2018.

[652] W. Cao, W. Song, X. Li, S. Zheng, G. Zhang, Y. Wu, S. He, H. Zhu, and J. Chen, "Interaction with social robots: Improving gaze toward face but not necessarily joint attention in children with autism spectrum disorder," *Frontiers in Psychology*, vol. 10, p. 1503, 2019.

[653] H.-L. Cao, R. E. Simut, N. Krepel, B. Vanderborght, and J. Vanderfaeillie, "Could NAO robot function as model demonstrating joint attention skills for children with autism spectrum disorder? an exploratory study," *International journal of humanoid robotics*, vol. 19, no. 04, p. 2240006, 2022.

[654] D. Casas-Bocanegra, D. Gomez-Vargas, M. J. Pinto-Bernal, J. Maldonado, M. Munera, A. Villa-Moreno, M. F. Stoelen, T. Belpaeme, and C. A. Cifuentes, "An open-source social robot based on compliant soft robotics for therapy with children with ASD," in *Actuators*, vol. 9, no. 3. MDPI, 2020, p. 91.

[655] E. Cervera, A. P. del Pobil, and M.-I. Cabezudo, "Playful interaction with humanoid robots for social development in autistic children: A pilot study," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–6.

[656] M. Chen, W. Xiao, L. Hu, Y. Ma, Y. Zhang, and G. Tao, "Cognitive wearable robotics for autism perception enhancement," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 4, pp. 1–16, 2021.

[657] P. Chevalier, J.-C. Martin, B. Isableu, C. Bazile, D.-O. Iacob, and A. Tapus, "Joint attention using human-robot interaction: Impact of sensory preferences of children with autism," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 849–854.

[658] P. Chevalier, G. Raiola, J.-C. Martin, B. Isableu, C. Bazile, and A. Tapus, "Do sensory preferences of children with autism impact an imitation task with a robot?" in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 177–186.

[659] P. Chevalier, D. Ghiglino, F. Floris, T. Priolo, and A. Wykowska, "Visual and hearing sensitivity affect robot-based training for children diagnosed with autism spectrum disorder," *Frontiers in Robotics and AI*, vol. 8, p. 748853, 2022.

[660] E. Y.-h. Chung, "Robotic intervention program for enhancement of social engagement among children with autism spectrum disorder," *Journal of Developmental and Physical Disabilities*, vol. 31, no. 4, pp. 419–434, 2019.

[661] D. Conti, S. Di Nuovo, S. Buono, G. Trubia, and A. Di Nuovo, "Use of robotics to stimulate imitation in children with autism spectrum disorder: A pilot study in a clinical setting," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2015, pp. 1–6.

[662] D. Conti, G. Trubia, S. Buono, S. Di Nuovo, and A. Di Nuovo, "Affect recognition in autism: A single case study on integrating a humanoid robot in a standard therapy," *QWERTY-Interdisciplinary Journal of Technology, Culture and Education*, vol. 14, no. 2, pp. 66–87, 2019.

[663] B. Coşkun, P. Uluer, E. Toprak, D. E. Barkana, H. Kose, T. Zorcec, B. Robins, and A. Landowska, "Stress detection of children with autism using physiological signals in Kaspar robot-based intervention studies," in *2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2022, pp. 01–07.

[664] S. Costa, J. Resende, F. O. Soares, M. J. Ferreira, C. P. Santos, and F. Moreira, "Applications of simple robots to encourage social receptiveness of adolescents with autism," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology society*. IEEE, 2009, pp. 5072–5075.

[665] C. A. Costescu, B. Vanderborght, and D. O. David, "Beliefs, emotions, and behaviors: Differences between children with ASD and typically developing children. a robot-enhanced task," *Journal of Evidence-Based Psychotherapies*, vol. 16, no. 2, 2016.

[666] ——, "Robot-enhanced CBT for dysfunctional emotions in social situations for children with ASD," *Journal of Evidence-Based Psychotherapies*, vol. 17, no. 2, 2017.

[667] D. O. David, C. A. Costescu, S. Matu, A. Szentagotai, and A. Dobrean, "Developing joint attention for children with autism in robot-enhanced therapy," *International Journal of Social Robotics (IJSR)*, vol. 10, no. 5, pp. 595–605, 2018.

[668] ——, "Effects of a robot-enhanced intervention for children with ASD on teaching turn-taking skills," *Journal of Educational Computing Research*, vol. 58, no. 1, pp. 29–62, 2020.

[669] P. S. Dehkordi, H. Moradi, M. Mahmoudi, and H. R. Pouretemad, "The design, development, and deployment of RoboParrot for screening autistic children," *International Journal of Social Robotics (IJSR)*, vol. 7, no. 4, pp. 513–522, 2015.

[670] L. Desideri, M. Negrini, M. C. Cutrone, A. Rouame, M. Malavasi, E.-J. Hoogerwerf, P. Bonifacci, R. Di Sarro *et al.*, "Exploring the use of a humanoid robot to engage children with autism spectrum disorder (ASD)." in *AAATE Conf.*, 2017, pp. 501–509.

[671] L. Desideri, M. Negrini, M. Malavasi, D. Tanzini, A. Rouame, M. C. Cutrone, P. Bonifacci, and E.-J. Hoogerwerf, "Using a humanoid robot as a complement to interventions for children with autism spectrum disorder: a pilot study," *Advances in Neurodevelopmental Disorders*, vol. 2, no. 3, pp. 273–285, 2018.

[672] M. Dimitrova, N. Vegt, and E. Barakova, "Designing a system of interactive robots for training collaborative skills to autistic children," in *2012 15th International Conference on Interactive Collaborative Learning (ICL)*. IEEE, 2012, pp. 1–8.

[673] R. I. Does, V. Nivedita, S. V. S. Reddy, D. L. F. Jana, M. Sudha, and R. R. Manjul, "Robot-assisted cognitive training for children with autism insights and outcomes," in *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2023, pp. 139–144.

[674] G. Ercolano, S. Rossi, D. Conti, and A. Di Nuovo, "Gesture recognition with a 2d low-resolution embedded camera to minimise intrusion in robot-led training of children with autism spectrum disorder," *Applied Intelligence*, vol. 54, no. 8, pp. 6579–6591, 2024.

[675] P. G. Esteban, P. Baxter, T. Belpaeme, E. Billing, H. Cai, H.-L. Cao, M. Coeckelbergh, C. Costescu, D. David, A. De Beir *et al.*, "How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder," *Paladyn, Journal of Behavioral Robotics*, vol. 8, no. 1, pp. 18–38, 2017.

[676] N. Fachantidis, C. K. Syriopoulou-Delli, I. Vezyrtzis, and M. Zygopoulou, "Beneficial effects of robot-mediated class activities on a child with ASD and his typical classmates," *International Journal of Developmental Disabilities*, vol. 66, no. 3, pp. 245–253, 2020.

[677] N. Fachantidis, C. K. Syriopoulou-Delli, and M. Zygopoulou, "The effectiveness of socially assistive robotics in children with autism spectrum disorder," *International journal of developmental disabilities*, vol. 66, no. 2, pp. 113–121, 2020.

[678] A. Febtriko, T. Rahayuningsih, D. Septiani, L. Trisnawati, D. Arisandi *et al.*, "Effectiveness of android-based mobile robots for children asperger syndrome," in *2018 International Conference on Applied Information Technology and Innovation (ICAITI)*. IEEE, 2018, pp. 208–212.

[679] D. Feil-Seifer and M. J. Mataric, "B 3 ia: A control architecture for autonomous robot-assisted behavior intervention for children with autism spectrum disorders," in *The 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2008, pp. 328–333.

[680] D. Feil-Seifer and M. Mataric, “Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features,” in *Proceedings of the 6th International Conference on Human-Robot Interaction (HRI)*, 2011, pp. 323–330.

[681] D. Feil-Seifer and M. J. Matarić, “Distance-based computational models for facilitating robot interaction with children,” *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 55–77, 2012.

[682] Y. Feng, Q. Jia, M. Chu, and W. Wei, “Engagement evaluation for autism intervention by robots based on dynamic bayesian network and expert elicitation,” *IEEE Access*, vol. 5, pp. 19 494–19 504, 2017.

[683] C. Fournier, C. Michelon, V. Granit, P. Audoyer, A. Bernardot, M.-C. Picot, A. Kheddar, and A. Baghdadli, “Pilot study protocol evaluating the impact of telerobotics interactions with autistic children during a denver intervention on communication skills using single-case experimental design,” *BMJ Open*, vol. 14, no. 11, p. e084110, 2024.

[684] D. François, S. Powell, and K. Dautenhahn, “A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children’s proactivity and initiative-taking,” *Interaction Studies*, vol. 10, no. 3, pp. 324–373, 2009.

[685] R. Fuentes-Alvarez, A. Morfin-Santana, K. Ibañez, I. Chairez, and S. Salazar, “Energetic optimization of an autonomous mobile socially assistive robot for autism spectrum disorder,” *Frontiers in Robotics and AI*, vol. 9, p. 1053115, 2023.

[686] J. Galán-Mena, G. Ávila, J. Pauta-Pintado, D. Lima-Juma, V. Robles-Bykbaev, and D. Quisi-Peralta, “An intelligent system based on ontologies and ICT tools to support the diagnosis and intervention of children with autism,” in *2016 IEEE Biennial Congress of Argentina (ARGENCON)*. IEEE, 2016, pp. 1–5.

[687] D. Ghiglino, P. Chevalier, F. Floris, T. Priolo, and A. Wykowska, “Follow the white robot: Efficacy of robot-assistive training for children with autism spectrum disorder,” *Research in Autism Spectrum Disorders*, vol. 86, p. 101822, 2021.

[688] I. Giannopulu, “Multimodal cognitive nonverbal and verbal interactions: the neurorehabilitation of autistic children via mobile toy robots,” *IARIA International Journal of Advances in Life Sciences*, vol. 5, no. 10.1109, 2013.

[689] J. Greczek, E. Kaszubski, A. Atrash, and M. Matarić, “Graded cueing feedback in robot-mediated imitation practice for children with autism spectrum disorders,” in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2014, pp. 561–566.

[690] M. Hirokawa, A. Funahashi, Y. Pan, Y. Itoh, and K. Suzuki, "Design of a robotic agent that measures smile and facing behavior of children with autism spectrum disorder," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 843–848.

[691] V. Holeva, V. Nikopoulou, C. Lytridis, C. Bazinas, P. Kechayas, G. Sidiropoulos, M. Papadopoulou, M. Kerasidou, C. Karatsioras, N. Geronikola *et al.*, "Effectiveness of a robot-assisted psychological intervention for children with autism spectrum disorder," *Journal of Autism and Developmental Disorders (JADD)*, vol. 54, no. 2, pp. 577–593, 2024.

[692] B. Ilijoski, N. Ackovska, T. Zorcec, and Z. Popeska, "Extending robot therapy for children with autism using mobile and web application," *Sensors*, vol. 22, no. 16, p. 5965, 2022.

[693] N. I. Ishak, H. M. Yusof, M. R. H. Ramlee, S. Na'im Sidek, and N. Rusli, "Modules of interaction for ASD children using RERO robot (humanoid)," in *2019 7th International Conference on Mechatronics Engineering (ICOM)*. IEEE, 2019, pp. 1–6.

[694] L. I. Ismail, S. Shamsudin, H. Yussof, F. A. Hanapiah, and N. I. Zahari, "Estimation of concentration by eye contact measurement in robot–based intervention program with autistic children," *Procedia Engineering*, vol. 41, pp. 1548–1552, 2012.

[695] ——, "Robot-based intervention program for autistic children with humanoid robot NAO: initial response in stereotyped behavior," *Procedia Engineering*, vol. 41, pp. 1441–1447, 2012.

[696] A. S. Ivani, A. Giubergia, L. Santos, A. Geminiani, S. Annunziata, A. Caglio, I. Olivieri, and A. Pedrocchi, "A gesture recognition algorithm in a robot therapy for ASD children," *Biomedical Signal Processing and Control*, vol. 74, p. 103512, 2022.

[697] H. Javed, M. Jeon, and C. H. Park, "Adaptive framework for emotional engagement in child-robot interactions for autism interventions," in *2018 15th International Conference on Ubiquitous Robots (UR)*. IEEE, 2018, pp. 396–400.

[698] H. Javed, W. Lee, and C. H. Park, "Toward an automated measure of social engagement for children with autism spectrum disorder—a personalized computational modeling approach," *Frontiers in Robotics and AI*, vol. 7, p. 43, 2020.

[699] K. Jordan, M. King, S. Hellersteth, A. Wirén, and H. Mulligan, "Feasibility of using a humanoid robot for enhancing attention and social skills in adolescents with autism spectrum disorder," *International journal of rehabilitation research*, vol. 36, no. 3, pp. 221–227, 2013.

[700] J. R. Kaboski, J. J. Diehl, J. Beriont, C. R. Crowell, M. Villano, K. Wier, and K. Tang, "Brief report: A pilot summer robotics camp to reduce social anxiety and improve social/vocational skills in adolescents with ASD," *Journal of Autism and Developmental Disorders (JADD)*, vol. 45, no. 12, pp. 3862–3869, 2015.

[701] E. Karakosta, K. Dautenhahn, D. S. Syrdal, L. J. Wood, and B. Robins, "Using the humanoid robot Kaspar in a greek school environment to support children with autism spectrum condition," *Paladyn, Journal of Behavioral Robotics*, vol. 10, no. 1, pp. 298–317, 2019.

[702] M. Karim, M. S. Mia, S. M. Tareeq, and M. Hasanuzzaman, "Evaluate effectiveness of NAO robot to train children with autism spectrum disorder (ASD)," in *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2023, pp. 165–174.

[703] M.-G. Kim, I. Oosterling, T. Lourens, W. Staal, J. Buitelaar, J. Glennon, I. Smeekens, and E. Barakova, "Designing robot-assisted pivotal response training in game activity for children with autism," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014, pp. 1101–1106.

[704] S. Kim, M. Hirokawa, S. Matsuda, A. Funahashi, and K. Suzuki, "Smiles as a signal of prosocial behaviors toward the robot in the therapeutic setting for children with autism spectrum disorder," *Frontiers in Robotics and AI*, vol. 8, p. 599755, 2021.

[705] S. Kim, M. Hirokawa, A. Funahashi, and K. Suzuki, "Parental influence in disengagement during robot-assisted activities: A case study of a parent and child with autism spectrum disorder," *Multimodal Technologies and Interaction*, vol. 6, no. 5, p. 39, 2022.

[706] S. Kim and Y. Iwai, "Students with autism spectrum disorder and intellectual disability teach a robot," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2024, pp. 1021–1028.

[707] S. A. Koch, C. E. Stevens, C. D. Clesi, J. B. Lebersfeld, A. G. Sellers, M. E. McNew, F. J. Biasini, F. R. Amthor, and M. I. Hopkins, "A feasibility study evaluating the emotionally expressive robot sam," *International Journal of Social Robotics (IJSR)*, vol. 9, no. 4, pp. 601–613, 2017.

[708] S. Konishi, M. Kuwata, Y. Matsumoto, Y. Yoshikawa, K. Takata, H. Haraguchi, A. Kudo, H. Ishiguro, and H. Kumazaki, "Self-administered questionnaires enhance emotion estimation of individuals with autism spectrum disorders in a robotic interview setting," *Frontiers in Psychiatry*, vol. 15, p. 1249000, 2024.

[709] J. Korneder, W.-Y. G. Louie, C. M. Pawluk, I. Abbas, M. Brys, and F. Rooney, "Robot-mediated interventions for teaching children with ASD: a new intraverbal skill," *Assistive Technology*, vol. 34, no. 6, pp. 707–716, 2022.

[710] M. W. De Korte, I. van den Berk-Smeekens, M. van Dongen-Boomsma, I. J. Oosterling, J. C. Den Boer, E. I. Barakova, T. Lourens, J. K. Buitelaar, J. C. Glennon, and W. G. Staal, "Self-initiations in young children with autism during pivotal response treatment with and without robot assistance," *Autism*, vol. 24, no. 8, pp. 2117–2128, 2020.

[711] V. Kostrubiec and J. Kruck, "Collaborative research project: Developing and testing a robot-assisted intervention for children with autism," *Frontiers in Robotics and AI*, vol. 7, p. 37, 2020.

[712] H. Kumazaki, Z. Warren, T. Muramatsu, Y. Yoshikawa, Y. Matsumoto, M. Miyao, M. Nakano, S. Mizushima, Y. Wakita, H. Ishiguro *et al.*, "A pilot study for robot appearance preferences among high-functioning individuals with autism spectrum disorder: Implications for therapeutic use," *PloS One*, vol. 12, no. 10, p. e0186581, 2017.

[713] H. Kumazaki, Y. Yoshikawa, Y. Yoshimura, T. Ikeda, C. Hasegawa, D. N. Saito, S. Tomiyama, K.-m. An, J. Shimaya, H. Ishiguro *et al.*, "The impact of robotic intervention on joint attention in children with autism spectrum disorders," *Molecular Autism*, vol. 9, no. 1, p. 46, 2018.

[714] H. Kumazaki, Z. Warren, A. Swanson, Y. Yoshikawa, Y. Matsumoto, Y. Yoshimura, J. Shimaya, H. Ishiguro, N. Sarkar, J. Wade *et al.*, "Brief report: Evaluating the utility of varied technological agents to elicit social attention from children with autism spectrum disorders," *Journal of Autism and Developmental Disorders (JADD)*, vol. 49, no. 4, pp. 1700–1708, 2019.

[715] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, H. Haraguchi, T. Sono, Y. Matsumoto, H. Ishiguro, M. Kikuchi, T. Sumiyoshi, and M. Mimura, "Enhancing communication skills of individuals with autism spectrum disorders while maintaining social distancing using two tele-operated robots," *Frontiers in Psychiatry*, vol. 11, p. 598688, 2021.

[716] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, Y. Matsumoto, M. Kuwata, K. Takata, H. Ishiguro, and M. Mimura, "Differences in the optimal motion of android robots for the ease of communications among individuals with autism spectrum disorders," *Frontiers in Psychiatry*, vol. 13, p. 883371, 2022.

[717] J. Y. Kwon, B. H. Lee, K.-H. Mun, and J. S. Jung, "A study on the effects of using an eco-friendly cat robot to treat children with autism spectrum disorder," *Indian Journal of Science and Technology*, vol. 8, no. 23, p. 1, 2015.

[718] G. Lakatos, L. J. Wood, D. S. Syrdal, B. Robins, A. Zaraki, and K. Dautenhahn, "Robot-mediated intervention can assist children with autism to develop visual

perspective taking skills," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 87–101, 2020.

[719] F. Lecciso, A. Levante, R. A. Fabio, T. Caprì, M. Leo, P. Carcagnì, C. Distante, P. L. Mazzeo, P. Spagnolo, and S. Petrocchi, "Emotional expression in children with ASD: a pre-study on a two-group pre-post-test design comparing robot-based and computer-based training," *Frontiers in Psychology*, vol. 12, p. 678052, 2021.

[720] J. Lee, H. Takehashi, C. Nagai, and G. Obinata, "Design of a therapeutic robot for interacting with autistic children through interpersonal touch," in *The 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2012, pp. 712–717.

[721] J. Lee, G. Obinata, D. Stefanov, and C. Nagai, "A preliminary study on the response of autistic persons to a robot feedback based on visual stimuli," *Paladyn, Journal of Behavioral Robotics*, vol. 4, no. 2, pp. 73–82, 2013.

[722] J. Lee and G. Obinata, "Developing therapeutic robot for children with autism: A study on exploring colour feedback," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 173–174.

[723] J. Lee, G. Obinata, and H. Aoki, "A pilot study of using touch sensing and robotic feedback for children with autism," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2014, pp. 222–223.

[724] J. Lin, J. Li, Y. She, L. Lin, H. Wu, E. Zhang, J. Lei, W. Huang, and J. Li, "Using a social robot for children with autism: A therapist-robot interactive model," *Computer Animation and Virtual Worlds*, vol. 33, no. 5, p. e2109, 2022.

[725] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Affect recognition in robot assisted rehabilitation of children with autism spectrum disorder," in *Proceedings 2007 IEEE International Conference on robotics and automation*. IEEE, 2007, pp. 1755–1760.

[726] ——, "Online affect detection and robot behavior adaptation for intervention of children with autism," *IEEE Transactions on Robotics*, vol. 24, no. 4, pp. 883–896, 2008.

[727] X. Liu, X. Zhou, C. Liu, J. Wang, X. Zhou, N. Xu, and A. Jiang, "An interactive training system of motor learning by imitation and speech instructions for children with autism," in *2016 9th International Conference on Human System Interactions (HSI)*. IEEE, 2016, pp. 56–61.

[728] G. Lorenzo and A. Lorenzo-Lledó, "The use of artificial intelligence for detecting the duration of autistic students' emotions in social interaction with the NAO robot: a case study," *International Journal of Information Technology*, vol. 16, no. 2, pp. 625–631, 2024.

[729] W.-Y. G. Louie, J. Korneder, I. Abbas, and C. Pawluk, “A study on an applied behavior analysis-based robot-mediated listening comprehension intervention for ASD,” *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 31–46, 2020.

[730] H. H. Lund, M. Dam Pedersen, and R. Beck, “Modular robotic tiles: experiments for children with autism,” *Artificial Life and Robotics*, vol. 13, no. 2, pp. 394–400, 2009.

[731] C. Lytridis, V. G. Kaburlasos, C. Bazinas, G. A. Papakostas, G. Sidiropoulos, V.-A. Nikopoulou, V. Holeva, M. Papadopoulou, and A. Evangeliou, “Behavioral data analysis of robot-assisted autism spectrum disorder (ASD) interventions based on lattice computing techniques,” *Sensors*, vol. 22, no. 2, p. 621, 2022.

[732] N. A. Malik, S. Shamsuddin, H. Yussof, M. A. Miskam, and A. C. Hamid, “Feasibility of using a humanoid robot to elicit communicational response in children with mild autism,” in *IOP Conference Series: Materials Science and Engineering*, vol. 53, no. 1. IOP Publishing, 2013, p. 012077.

[733] M. D. Manner, “Using small humanoid robots to detect autism in toddlers.” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015, pp. 4383–4384.

[734] N. Martínez, V. López, J. Albo-Canals, O. Sans, and L. Toppin, “Implementation of a robotic assistant to develop and monitor learning activities in children with autism,” in *2022 8th International Engineering, Sciences and Technology Conference (IESTEC)*. IEEE, 2022, pp. 678–683.

[735] S. M. Mavadati, H. Feng, A. Gutierrez, and M. H. Mahoor, “Comparing the gaze responses of children with autism and typically developed individuals in human-robot interaction,” in *2014 IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2014, pp. 1128–1133.

[736] S. M. Mavadati, H. Feng, M. Salvador, S. Silver, A. Gutierrez, and M. H. Mahoor, “Robot-based therapeutic protocol for training children with autism,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 855–860.

[737] M. Mayadunne, U. A. Manawadu, K. R. Abeyratne, and P. R. S. De Silva, “A robotic companion for children diagnosed with autism spectrum disorder,” in *2020 International Conference on Image Processing and Robotics (ICIP)*. IEEE, 2020, pp. 1–6.

[738] D. Mazzei, L. Billeci, A. Armato, N. Lazzeri, A. Cisternino, G. Pioggia, R. Igliozzi, F. Muratori, A. Ahluwalia, and D. De Rossi, “The face of autism,” in *19th International Symposium in Robot and Human Interactive Communication*. IEEE, 2010, pp. 791–796.

[739] D. Mazzei, A. Greco, N. Lazzeri, A. Zaraki, A. Lanata, R. Igliozzi, A. Mancini, F. Stoppa, E. P. Scilingo, F. Muratori *et al.*, "Robotic social therapy on children with autism: preliminary evaluation through multi-parametric analysis," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing.* IEEE, 2012, pp. 766–771.

[740] F. Mehmood, H. Mahzoon, Y. Yoshikawa, H. Ishiguro, H. Sadia, S. Ali, and Y. Ayaz, "Attentional behavior of children with ASD in response to robotic agents," *IEEE Access*, vol. 9, pp. 31 946–31 955, 2021.

[741] B. Mehralizadeh, P. Soleiman, S. Nikkhoo, M. Rahimi, A. Kargar, F. Masoumi, and H. Moradi, "Multi-modal ASD screening system: A preliminary study," in *2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM).* IEEE, 2023, pp. 228–234.

[742] A. N. Meltzoff, R. Brooks, A. P. Shon, and R. P. Rao, "'social'"' robots are psychological agents for infants: A test of gaze following," *Neural Networks*, vol. 23, no. 8-9, pp. 966–972, 2010.

[743] S. E. Mengoni, K. Irvine, D. Thakur, G. Barton, K. Dautenhahn, K. Guldberg, B. Robins, D. Wellsted, and S. Sharma, "Feasibility study of a randomised controlled trial to investigate the effectiveness of using a humanoid robot to improve the social skills of children with autism spectrum disorder (Kaspar RCT): A study protocol," *BMJ Open*, vol. 7, no. 6, p. e017376, 2017.

[744] R. S. Moorthy and S. Pugazhenthi, "Imitation based training to enhance psychomotor skills in autistic children using a snatcher robot," in *2016 International Conference on Robotics and Automation for Hhumanitarian Applications (RAHA).* IEEE, 2016, pp. 1–6.

[745] K. Niderla and M. Maciejewski, "Construction of a zoomorphical robot for rehabilitation of autistic children," in *Journal of Physics: Conference Series*, vol. 1782, no. 1. IOP Publishing, 2021, p. 012024.

[746] G. Nie, Z. Zheng, J. Johnson, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Predicting response to joint attention performance in human-human interaction based on human-robot interaction for young children with autism spectrum disorder," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2018, pp. 1–4.

[747] G. Nie, Z. Zheng, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "An exploration into the design of multi-session robot-mediated joint attention intervention for young children with autism," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2024, pp. 1836–1841.

[748] E. Nunez, S. Matsuda, M. Hirokawa, J. Yamamoto, and K. Suzuki, “Paired robotic devices to mediate and represent social behaviors,” in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2015, pp. 722–727.

[749] A. Di Nuovo, D. Conti, G. Trubia, S. Buono, and S. Di Nuovo, “Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability,” *Robotics*, vol. 7, no. 2, p. 25, 2018.

[750] J. Oliver, R. Oliván, J. Shukla, A. Folch, R. Martínez-Leal, M. Castellá, and D. Puig, “Stakeholders acceptance and expectations of robot-assisted therapy for children with autism spectrum disorder,” in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2019, pp. 1–7.

[751] M. T. van Otterdijk, M. W. de Korte, I. van den Berk-Smeekens, J. Hendrix, M. van Dongen-Boomsma, J. C. Den Boer, J. K. Buitelaar, T. Lourens, J. C. Glennon, W. G. Staal *et al.*, “The effects of long-term child–robot interaction on the attention and the engagement of children with autism,” *Robotics*, vol. 9, no. 4, p. 79, 2020.

[752] C. Paengkumhag, W. Jutharee, W. Limpornchitwilai, K. Chamnongthai, and B. Keawkamnerdpong, “Using robot-assisting personalized learning for children with autism: A pilot study of robot’s actions,” in *2023 15th Biomedical Engineering International Conference (BMEiCON).* IEEE, 2023, pp. 1–5.

[753] R. Pakkar, C. Clabaugh, R. Lee, E. Deng, and M. J. Matarićć, “Designing a socially assistive robot for long-term in-home use for children with autism spectrum disorders,” in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2019, pp. 1–7.

[754] G. Palestra, B. De Carolis, F. Esposito *et al.*, “Artificial intelligence for robot-assisted treatment of autism.” in *Waiah@ ai* ia*, 2017, pp. 17–24.

[755] G. Palestra, D. Cazzato, F. Adamo, I. Bortone, and C. Distante, “Assistive robot, RGB-D sensor and graphical user interface to encourage communication skills in ASD population,” *Journal of Medical Robotics Research*, vol. 2, no. 02, p. 1740002, 2017.

[756] J. A. C. Panceri, É. Freitas, J. C. de Souza, S. da Luz Schreider, E. Caldeira, and T. F. Bastos, “A new socially assistive robot with integrated serious games for therapies with children with autism spectrum disorder and down syndrome: a pilot study,” *Sensors*, vol. 21, no. 24, p. 8414, 2021.

[757] T. Papazoglou, C. Karagiannidis, and S. Mavropoulou, “Educational robotics can foster social inclusion and social status of children with autism,” in *2021 International Conference on Advanced Learning Technologies (ICALT).* IEEE, 2021, pp. 317–319.

[758] A. Peca, R. Simut, S. Pintea, C. Costescu, and B. Vanderborght, "How do typically developing children and children with autism perceive different social robots?" *Computers in Human Behavior*, vol. 41, pp. 268–277, 2014.

[759] E. Pérez-Vázquez, G. L. Lledó, A. L. Carreres, A. Lorenzo-Lledó, and A. Gilabert-Cerdá, "Application of the Bee-Bot robot to improve the oral expression of a student with ASD: a case study," in *2023 XIII International Conference on Virtual Campus (JICV).* IEEE, 2023, pp. 1–4.

[760] E. Peribañez, S. Bayona, J. San Martin, A. Verde, C. Garre, J. Leoste, and L. Pastor, "An experimental methodology for introducing educational robotics and storytelling in therapeutical activities for children with neurodevelopmental disorders," *Machines*, vol. 11, no. 6, p. 629, 2023.

[761] F. Petric, D. Miklić, M. Cepanec, P. Cvitanović, and Z. Kovačić, "Functional imitation task in the context of robot-assisted autism spectrum disorder diagnostics: Preliminary investigations," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, 2017, pp. 1471–1478.

[762] A. C. Pierno, M. Mari, D. Lusher, and U. Castiello, "Robotic movement elicits visuomotor priming in children with autism," *Neuropsychologia*, vol. 46, no. 2, pp. 448–454, 2008.

[763] M. J. Pinto-Bernal, N. Cespedes, P. Castro, M. Munera, and C. A. Cifuentes, "Physical human-robot interaction influence in ASD therapy through an affordable soft social robot," *Journal of Intelligent & Robotic Systems*, vol. 105, no. 3, p. 67, 2022.

[764] G. Pioggia, R. Igliozzi, M. Ferro, A. Ahluwalia, F. Muratori, and D. De Rossi, "An android for enhancing social skills and emotion recognition in people with autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 4, pp. 507–515, 2005.

[765] G. Tartarisco, R. Bruschetta, F. Marino, T. Caprì, R. Minutoli, P. Chilà, C. Failla, A. Puglisi, A. A. Arnao, A. Cerasa *et al.*, "Exploring behavioural and physiological interactions in a group-based emotional skill social robotic training for autism spectrum disorders," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME).* IEEE, 2022, pp. 1–4.

[766] C. A. Pop, A. C. Petrule, S. Pintea, A. Peca, R. Simut, B. Vanderborght, and D. O. David, "Imitation and social behaviors of children with ASD in interaction with robonova. a series of single case experiments." *Transylvanian Journal of Psychology/Erdélyi Pszichológiai Szemle*, vol. 14, no. 1, 2013.

[767] C. A. Pop, R. E. Simut, S. Pintea, J. Saldien, A. S. Rusu, J. Vanderfaeillie, D. O. David, D. Lefeber, and B. Vanderborght, "Social robots vs. computer

display: Does the way social stories are delivered make a difference for their effectiveness on ASD children?” *Journal of Educational Computing Research*, vol. 49, no. 3, pp. 381–401, 2013.

[768] C. A. Pop, S. Pintea, B. Vanderborght, and D. O. David, “Enhancing play skills, engagement and social skills in a play task in ASD children by using robot-based interventions. a pilot study,” *Interaction Studies*, vol. 15, no. 2, pp. 292–320, 2014.

[769] A. Ghorbandaei Pour, A. Taheri, M. Alemi, and A. Meghdari, “Human–robot facial expression reciprocal interaction platform: case studies on children with autism,” *International Journal of Social Robotics (IJSR)*, vol. 10, no. 2, pp. 179–198, 2018.

[770] G. Pradel, P. Dansart, A. Puret, and C. Barthélemy, “Generating interactions in autistic spectrum disorders by means of a mobile robot,” in *IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society*. IEEE, 2010, pp. 1540–1545.

[771] M. Puyon and I. Giannopulu, “Emergent emotional and verbal strategies in autism are based on multimodal interactions with toy robots in free spontaneous game play,” in *The 22nd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2013, pp. 593–597.

[772] U. Qidwai, S. B. A. Kashem, and O. Conor, “Humanoid robot as a teacher’s assistant: Helping children with autism to learn social and academic skills,” *Journal of Intelligent & Robotic Systems*, vol. 98, no. 3, pp. 759–770, 2020.

[773] N. Rakhymbayeva, A. Amirova, and A. Sandygulova, “A long-term engagement with a social robot for autism therapy,” *Frontiers in Robotics and AI*, vol. 8, p. 669972, 2021.

[774] A. A. Ramirez-Duque, A. Frizera-Neto, and T. F. Bastos, “Robot-assisted diagnosis for children with autism spectrum disorder based on automated analysis of nonverbal cues,” in *2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob)*. IEEE, 2018, pp. 456–461.

[775] ——, “Robot-assisted autism spectrum disorder diagnostic based on artificial reasoning,” *Journal of Intelligent & Robotic Systems*, vol. 96, no. 2, pp. 267–281, 2019.

[776] A. A. Ramirez-Duque, T. Bastos, M. Munera, C. A. Cifuentes, and A. Frizera-Neto, “Robot-assisted intervention for children with special needs: A comparative assessment for autism screening,” *Robotics and Autonomous Systems*, vol. 127, p. 103484, 2020.

[777] V. Robles-Bykbaev, J. Galán-Mena, M. López-Nores, Y. Robles-Bykbaev, J. Cornejo-Reyes, D. Quisi-Peralta, J. Pazos-Arias, D. Lima-Juma, and V. C.

León-Wong, “A hybrid system to support the psychological intervention for children with autism: A proposal based on knowledge modeling and intelligent technologies,” *International Journal of E-Services and Mobile Applications (IJESMA)*, vol. 10, no. 4, pp. 74–100, 2018.

[778] Á. Rodríguez-Quevedo, G. L. Lledó, A. L. Carreres, and E. Pérez-Vázquez, “Emotion recognition in ASD students with social robotics: Case study,” in *2023 XIII International Conference on Virtual Campus (JICV)*. IEEE, 2023, pp. 1–4.

[779] R. Romero-García, R. Martínez-Tomás, P. Pozo, F. de la Paz, and E. Sarriá, “Q-CHAT-NAO: A robotic approach to autism screening in toddlers,” *Journal of Biomedical Informatics*, vol. 118, p. 103797, 2021.

[780] O. Rudovic, J. Lee, L. Mascarell-Maricic, B. W. Schuller, and R. W. Picard, “Measuring engagement in robot-assisted autism therapy: A cross-cultural study,” *Frontiers in Robotics and AI*, vol. 4, p. 36, 2017.

[781] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, “Personalized machine learning for robot perception of affect and engagement in autism therapy,” *Science Robotics*, vol. 3, no. 19, p. eaao6760, 2018.

[782] M. N. Saadatzi, R. C. Pennington, K. C. Welch, and J. H. Graham, “Effects of a robot peer on the acquisition and observational learning of sight words in young adults with autism spectrum disorder,” *Journal of Special Education Technology*, vol. 33, no. 4, pp. 284–296, 2018.

[783] ——, “Small-group technology-assisted instruction: Virtual teacher and robot peer for individuals with autism spectrum disorder,” *Journal of Autism and Developmental Disorders (JADD)*, vol. 48, no. 11, pp. 3816–3830, 2018.

[784] P. Saha, M. I. Tapotee, and M. A. R. Ahad, “Task detection of ASD children by analyzing robotic enhanced and standard human therapy,” in *2021 Thirteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*. IEEE, 2021, pp. 1–6.

[785] T. Salter, K. Dautenhahn, and R. Te Boekhorst, “Learning about natural human–robot interaction styles,” *Robotics and Autonomous Systems*, vol. 54, no. 2, pp. 127–134, 2006.

[786] M. J. Salvador, S. Silver, and M. H. Mahoor, “An emotion recognition comparative study of autistic and typically-developing children using the zeno robot,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 6128–6133.

[787] A. Sandygulova, Z. Zhexenova, B. Tleubayev, A. Nurakhmetova, D. Zhumabekova, I. Assylgali, Y. Rzagaliyev, and A. Zhakenova, “Interaction design and methodology of robot-assisted therapy for children with severe ASD and

ADHD," *Paladyn, Journal of Behavioral Robotics*, vol. 10, no. 1, pp. 330–345, 2019.

[788] A. Sandygulova, A. Amirova, Z. Telisheva, A. Zhanatkyzy, and N. Rakhymbayeva, "Individual differences of children with autism in robot-assisted autism therapy," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 43–52.

[789] L. Santos, A. Geminiani, I. Olivieri, J. Santos-Victor, and A. Pedrocchi, "Copyrobot: Interactive mirroring robotics game for ASD children," in *Mediterranean Conference on Medical and Biological Engineering and Computing*. Springer, 2019, pp. 2014–2027.

[790] L. Santos, B. Silva, F. Maddaloni, A. Geminiani, A. Caglio, S. Annunziata, I. Olivieri, C. Barata, J. Santos-Victor, and A. Pedrocchi, "Sharing worlds: design of a real-time attention classifier for robotic therapy of ASD children," in *2022 International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2022, pp. 1–6.

[791] B. R. Schadenberg, D. Reidsma, D. K. Heylen, and V. Evers, "Differences in spontaneous interactions of autistic children in an interaction with an adult and humanoid robot," *Frontiers in Robotics and AI*, vol. 7, p. 28, 2020.

[792] B. R. Schadenberg, D. Reidsma, V. Evers, D. P. Davison, J. J. Li, D. K. Heylen, C. Neves, P. Alvito, J. Shen, M. Pantić *et al.*, "Predictable robots for autistic children—variance in robot behaviour, idiosyncrasies in autistic children's characteristics, and child–robot engagement," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 28, no. 5, pp. 1–42, 2021.

[793] S. Shamsuddin, H. Yussof, L. I. Ismail, S. Mohamed, F. A. Hanapiah, and N. I. Zahari, "Humanoid robot NAO interacting with autistic children of moderately impaired intelligence to augment communication skills," *Procedia Engineering*, vol. 41, pp. 1533–1538, 2012.

[794] ——, "Initial response in HRI-a case study on evaluation of child with autism spectrum disorders interacting with a humanoid robot Nao," *Procedia Engineering*, vol. 41, pp. 1448–1455, 2012.

[795] S. Shamsuddin, H. Yussof, F. A. Hanapiah, and S. Mohamed, "A qualitative method to analyze response in robotic intervention for children with autism," in *The 22nd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2013, pp. 324–325.

[796] ——, "Response of children with autism to robotic intervention and association with IQ levels," in *4th International Conference on Development and Learning and on Epigenetic Robotics*. IEEE, 2014, pp. 387–393.

[797] T. She, X. Kang, S. Nishide, and F. Ren, "Improving leo robot conversational ability via deep learning algorithms for children with autism," in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*. IEEE, 2018, pp. 416–420.

[798] T. She and F. Ren, "Enhance the language ability of humanoid robot NAO through deep learning to interact with autistic children," *Electronics*, vol. 10, no. 19, p. 2393, 2021.

[799] Z. Shi, T. R. Groechel, S. Jain, K. Chima, O. Rudovic, and M. J. Matarić, "Toward personalized affect-aware socially assistive robot tutors for long-term interventions with children with autism," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 4, pp. 1–28, 2022.

[800] P. R. S. De Silva, K. Tadano, A. Saito, S. G. Lambacher, and M. Higashi, "The development of an assistive robot for improving the joint attention of autistic children," in *2009 IEEE International Conference on rehabilitation robotics*. IEEE, 2009, pp. 694–700.

[801] P. Ravindra, S. De Silva, K. Tadano, A. Saito, S. G. Lambacher, and M. Higashi, "Therapeutic-assisted robot for children with autism," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 3561–3567.

[802] B. Silva, L. Santos, C. Barata, A. Geminiani, G. Fassina, A. R. Gonzalez, S. Ferreira, B. Barahona-Corrêa, I. Olivieri, A. Pedrocchi *et al.*, "Attention analysis in robotic-assistive therapy for children with autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 2220–2229, 2024.

[803] D. Silvera-Tawil, D. Bradford, and C. Roberts-Yates, "Talk to me: The role of human-robot interaction in improving verbal communication skills in students with autism or intellectual disability," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 1–6.

[804] R. E. Simut, J. Vanderfaeillie, A. Peca, G. Van de Perre, and B. Vanderborght, "Children with autism spectrum disorders make a fruit salad with Probo, the social robot: an interaction study," *Journal of Autism and Developmental Disorders (JADD)*, vol. 46, no. 1, pp. 113–126, 2016.

[805] R. Simut, G. Van de Perre, C. Costescu, J. Saldien, J. Vanderfaeillie, D. David, D. Lebefer, and B. Vanderborght, "Probogotchi: A novel edutainment device as a bridge for interaction between a child with ASD and the typically developed sibling," *Journal of Evidence-Based Psychotherapies*, vol. 16, no. 1, pp. 91–112, 2016.

[806] A. Singh, K. Raj, T. Kumar, S. Verma, and A. M. Roy, "Deep learning-based cost-effective and responsive robot for autism treatment," *Drones*, vol. 7, no. 2, p. 81, 2023.

[807] W.-C. So, M. K.-Y. Wong, W.-Y. Lam, C.-H. Cheng, J.-H. Yang, Y. Huang, P. Ng, W.-L. Wong, C.-L. Ho, K.-L. Yeung *et al.*, "Robot-based intervention may reduce delay in the production of intransitive gestures in chinese-speaking preschoolers with autism spectrum disorder," *Molecular Autism*, vol. 9, no. 1, p. 34, 2018.

[808] W.-C. So, M. K.-Y. Wong, C. K.-Y. Lam, W.-Y. Lam, A. T.-F. Chui, T.-L. Lee, H.-M. Ng, C.-H. Chan, and D. C.-W. Fok, "Using a social robot to teach gestural recognition and production in children with autism spectrum disorders," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 6, pp. 527–539, 2018.

[809] W.-C. So, C.-H. Cheng, W.-Y. Lam, T. Wong, W.-W. Law, Y. Huang, K.-C. Ng, H.-C. Tung, and W. Wong, "Robot-based play-drama intervention may improve the narrative abilities of chinese-speaking preschoolers with autism spectrum disorder," *Research in Developmental Disabilities*, vol. 95, p. 103515, 2019.

[810] W.-C. So, M. K.-Y. Wong, W.-Y. Lam, C.-H. Cheng, S.-Y. Ku, K.-Y. Lam, Y. Huang, and W.-L. Wong, "Who is a better teacher for children with autism? comparison of learning outcomes between robot-based and human-based interventions in gestural production and recognition," *Research in Developmental Disabilities*, vol. 86, pp. 62–75, 2019.

[811] W.-C. So, C.-H. Cheng, W.-W. Law, T. Wong, C. Lee, F.-Y. Kwok, S.-H. Lee, and K.-Y. Lam, "Robot dramas may improve joint attention of chinese-speaking low-functioning children with autism: Stepped wedge trials," *Disability and Rehabilitation: Assistive Technology*, vol. 18, no. 2, pp. 195–204, 2023.

[812] W.-C. So, W.-W. Law, C.-H. Cheng, C. Lee, K.-C. Ng, F.-Y. Kwok, H.-W. Lam, and K.-Y. Lam, "Comparing the effectiveness of robot-based to human-based intervention in improving joint attention in autistic children," *Frontiers in Psychiatry*, vol. 14, p. 1114907, 2023.

[813] F. O. Soares, S. C. Costa, C. P. Santos, A. P. S. Pereira, A. R. Hiolle, and V. Silva, "Socio-emotional development in high functioning children with autism spectrum disorders using a humanoid robot," *Interaction Studies*, vol. 20, no. 2, pp. 205–233, 2019.

[814] V. Sperati, B. Özcan, L. Romano, T. Moretta, S. Scaffaro, N. Faedda, G. Turturo, F. Fioriello, S. Pelosi, F. Giovannone *et al.*, "Acceptability of the transitional wearable companion "+ me" in children with autism spectrum disorder: A comparative pilot study," *Frontiers in Psychology*, vol. 11, p. 951, 2020.

[815] S. M. Srinivasan, I.-M. Eigsti, L. Neelly, and A. N. Bhat, "The effects of embodied rhythm and robotic interventions on the spontaneous and responsive social attention patterns of children with autism spectrum disorder (ASD): A pilot randomized controlled trial," *Research in Autism Spectrum Disorders*, vol. 27, pp. 54–72, 2016.

[816] C. M. Stanton, P. H. Kahn Jr, R. L. Severson, J. H. Ruckert, and B. T. Gill, "Robotic animals might aid in the social development of children with autism," in *Proceedings of the 3rd ACM/IEEE International Conference on Human robot interaction*, 2008, pp. 271–278.

[817] C. L. van Straten, I. Smeekens, E. Barakova, J. Glennon, J. Buitelaar, and A. Chen, "Effects of robots' intonation and bodily appearance on robot-mediated communicative treatment outcomes for children with autism spectrum disorder," *Personal and Ubiquitous Computing*, vol. 22, no. 2, pp. 379–390, 2018.

[818] R. Suzuki and J. Lee, "Robot-play therapy for improving prosocial behaviours in children with autism spectrum disorders," in *2016 International Symposium on Micro-NanoMechatronics and Human Science (MHS)*. IEEE, 2016, pp. 1–5.

[819] D. S. Syrdal, K. Dautenhahn, B. Robins, E. Karakosta, and N. C. Jones, "Kaspar in the wild: Experiences from deploying a small humanoid robot in a nursery school for children with autism," *Paladyn, Journal of Behavioral Robotics*, vol. 11, no. 1, pp. 301–326, 2020.

[820] A. Taheri, A. Meghdari, M. Alemi, and H. Pouretemad, "Clinical interventions of social humanoid robots in the treatment of a set of high-and low-functioning autistic Iranian twins," *Scientia Iranica*, vol. 25, no. 3, pp. 1197–1214, 2018.

[821] ——, "Human–robot interaction in autism treatment: a case study on three pairs of autistic children as twins, siblings, and classmates," *International Journal of Social Robotics (IJSR)*, vol. 10, no. 1, pp. 93–113, 2018.

[822] A. Taheri, A. Shariati, R. Heidari, M. Shahab, M. Alemi, and A. Meghdari, "Impacts of using a social robot to teach music to children with low-functioning autism," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 256–275, 2021.

[823] A. Talaei-Khoei, L. Lewis, M. Kaul, J. Daniel, and R. Sharma, "Use of lean robotic communication to improve social response of children with autism," in *AMCIS 2017-America's Conference on Information Systems: A Tradition of Innovation*, 2017.

[824] Z. Telisheva, A. Amirova, N. Rakhymbayeva, A. Zhanatkyzy, and A. Sandygulova, "The quantitative case-by-case analyses of the socio-emotional outcomes of children with ASD in robot-assisted autism therapy," *Multimodal Technologies and Interaction*, vol. 6, no. 6, p. 46, 2022.

[825] B. Tleubayev, Z. Zhexenova, A. Zhakenova, and A. Sandygulova, "Robot-assisted therapy for children with ADHD and ASD: a pilot study," in *Proceedings of the 2019 2nd International Conference on Service Robotics Technologies*, 2019, pp. 58–62.

[826] C. T. Valadão, C. Goulart, H. Rivera, E. Caldeira, T. F. Bastos Filho, A. Frizera-Neto, and R. Carelli, "Analysis of the use of a robot to improve social skills in children with autism spectrum disorder," *Research on Biomedical Engineering*, vol. 32, no. 2, pp. 161–175, 2016.

[827] B. Vanderborght, R. Simut, J. Saldien, C. Pop, A. S. Rusu, S. Pintea, D. Lefeber, and D. O. David, "Using the social robot Probo as a social story telling agent for children with ASD," *Interaction Studies*, vol. 13, no. 3, pp. 348–372, 2012.

[828] M. Villano, C. R. Crowell, K. Wier, K. Tang, B. Thomas, N. Shea, L. M. Schmitt, and J. J. Diehl, "DOMER: a wizard of oz interface for using interactive robots to scaffold social skills for children with autism spectrum disorders," in *Proceedings of the 6th International Conference on Human-Robot Interaction (HRI)*, 2011, pp. 279–280.

[829] J. Wainer, K. Dautenhahn, B. Robins, and F. Amirabdollahian, "Collaborating with Kaspar: Using an autonomous humanoid robot to foster cooperative dyadic play among children with autism," in *2010 10th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2010, pp. 631–638.

[830] G.-b. Wan, F.-h. Deng, Z.-j. Jiang, S.-z. Lin, C.-l. Zhao, B.-x. Liu, G. Chen, S.-h. Chen, X.-h. Cai, H.-b. Wang *et al.*, "Attention shifting during child—robot interaction: A preliminary clinical study for children with autism spectrum disorder," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 3, pp. 374–387, 2019.

[831] P. Wanglavan, W. Jutharee, T. Maneewarn, and B. Kaewkamnerdpong, "The development of attention detection model from child behavior for robot-assisted autism therapy," in *2019 19th International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2019, pp. 775–780.

[832] Z. Warren, Z. Zheng, S. Das, E. M. Young, A. Swanson, A. Weitlauf, and N. Sarkar, "Brief report: development of a robotic intervention platform for young children with ASD," *Journal of Autism and Developmental Disorders (JADD)*, vol. 45, no. 12, pp. 3870–3876, 2015.

[833] Z. E. Warren, Z. Zheng, A. R. Swanson, E. Bekele, L. Zhang, J. A. Crittendon, A. F. Weitlauf, and N. Sarkar, "Can robotic interaction improve joint attention skills?" *Journal of Autism and Developmental Disorders (JADD)*, vol. 45, no. 11, pp. 3726–3734, 2015.

[834] K. C. Welch, R. Pennington, S. Vanaparthy, H. M. Do, R. Narayanan, D. Popa, G. Barnes, and G. Kuravackel, "Using physiological signals and machine learning algorithms to measure attentiveness during robot-assisted social skills intervention: A case study of two children with autism spectrum disorder," *IEEE Instrumentation & Measurement Magazine*, vol. 26, no. 3, pp. 39–45, 2023.

[835] H. Wong and Z. Zhong, "Assessment of robot training for social cognitive learning," in *2016 16th International Conference on control, automation and systems (ICCAS)*. IEEE, 2016, pp. 893–898.

[836] R. Cañete Yaque, H. Svarrer Larsen, and M. E. Peralta Alvarez, "Pepe: an adaptive robot that helps children with autism to plan and self-manage their day," in *Proceedings of the 11th International Conference on the Internet of Things*, 2021, pp. 223–227.

[837] Y. Yoshikawa, H. Kumazaki, Y. Matsumoto, M. Miyao, M. Kikuchi, and H. Ishiguro, "Relaxing gaze aversion of adolescents with autism spectrum disorder in consecutive conversations with human and android robot—a preliminary study," *Frontiers in Psychiatry*, vol. 10, p. 370, 2019.

[838] S.-S. Yun, S.-K. Park, and J. Choi, "A robotic treatment approach to promote social interaction skills for children with autism spectrum disorders," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2014, pp. 130–134.

[839] S.-S. Yun, H. Kim, J. Choi, and S.-K. Park, "A robot-assisted behavioral intervention system for children with autism spectrum disorders," *Robotics and Autonomous Systems*, vol. 76, pp. 58–67, 2016.

[840] S.-S. Yun, J. Choi, and S.-K. Park, "Robotic behavioral intervention to facilitate eye contact and reading emotions of children with autism spectrum disorders," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 694–699.

[841] H. Yussof, S. Shamsuddin, F. A. Hanapiah, L. I. Ismail, and M. A. Miskam, "IQ level assessment methodology in robotic intervention with children with autism," in *2015 10TH Asian Control Conference (ASCC)*. IEEE, 2015, pp. 1–6.

[842] A. Zaraki, L. Wood, B. Robins, and K. Dautenhahn, "Development of a semi-autonomous robotic system to assist children with autism in developing visual perspective taking skills," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 969–976.

[843] Y. Zhang, W. Song, Z. Tan, H. Zhu, Y. Wang, C. M. Lam, Y. Weng, S. P. Hoi, H. Lu, B. S. M. Chan *et al.*, "Could social robots facilitate children with autism spectrum disorders in learning distrust and deception?" *Computers in Human Behavior*, vol. 98, pp. 140–149, 2019.

[844] Y. Zhang, W. Song, Z. Tan, Y. Wang, C. M. Lam, S. P. Hoi, Q. Xiong, J. Chen, and L. Yi, "Theory of robot mind: False belief attribution to social robots in children with and without autism," *Frontiers in Psychology*, vol. 10, p. 1732, 2019.

[845] Z. Zheng, L. Zhang, E. Bekele, A. Swanson, J. A. Crittendon, Z. Warren, and N. Sarkar, "Impact of robot-mediated interaction system on joint attention skills for children with autism," in *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2013, pp. 1–8.

[846] Z. Zheng, S. Das, E. M. Young, A. Swanson, Z. Warren, and N. Sarkar, "Autonomous robot-mediated imitation learning for children with autism," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 2707–2712.

[847] Z. Zheng, E. M. Young, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Robot-mediated mixed gesture imitation skill training for young children with ASD," in *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 72–77.

[848] Z. Zheng, H. Zhao, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Design, development, and evaluation of a noninvasive autonomous robot-mediated joint attention intervention system for young children with ASD," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 125–135, 2017.

Appendix A: Review Corpus and Summary Table

The table below presents a compilation of data and key characteristics extracted from the studies reviewed in Chapter 2. Each entry reflects the specific information used in our analysis, including: study duration (in days), number of user sessions, number of participants with complete data, type of HRI interaction, study domain, robot platform, deployment location, participant age group, type of results reported, and the level of robot autonomy. To improve readability, the following abbreviations are used: *unk.* = unknown or not reported by the study authors; Ql = qualitative data; Qn = quantitative data; A = fully autonomous; S = semi-autonomous; N = non-autonomous. For space considerations, study domains are abbreviated in the table but correspond to the categories outlined in Section 2.4.2.

	Per User									
Citation	**Period (Days)**	**# Sessions**	**Participants**	**HRI Type**	**Domain**	**Robot Used**	**Location**	**Age Group**	**Results**	**Operation**
[51] Afyouni et al., 2022	7	*unk.*	8	Diadic	Physical Health	Pepper	Facility	Adults	Ql	A
[100] Agrigoroaie & Tapus, 2018	4	8	1	Diadic	Mental Health	Tiago	Home	Elderly	Qn	A
[105] Ahmad et al., 2017	10	9	23	Diadic	General Purpose	NAO	School	Children	Ql, Qn	N
[81] Ahtinen et al., 2023	28	*unk.*	32	Family	Education	Alpha Mini	Home	Children	Ql	S
[163] Alemi et al., 2014	35	10	30	Group	Education	NAO	School	Children	Qn	A
[41] Bajones et al., 2019	21	free	16	Diadic	Physical Health	Hobbit PT2	Home	Elderly	Ql, Qn	A
[43] Barco et al., 2014	180	120	15	Diadic	Mental Health	LEGO	Home	Mixed	Qn	A
[113] Baxter et al., 2017	14	3	59	Diadic	Education	NAO	School	Children	Qn	A
[147] Begum et al., 2015	8	18	3	Triadic	ASD	NAO	School	Children	Ql, Qn	N
[53] Bodala et al., 2021	35	5	9	Group	Mental Health	Pepper	Lab	Mixed	Ql, Qn	A
[68] Cagiltay et al., 2022	28	12	14	Diadic	Education	Misty	Home	Children	Ql	A
[154] Carrillo et al., 2018	14	14	9	Diadic	Physical Health	NAO	Hospital	Children	Ql, Qn	A
[130] Chandra et al., 2018	30	4	25	Diadic	Education	NAO	School	Children	Qn	A
[67] Chen et al., 2022	32	6	12	Triadic	Education	Jibo	Home	Mixed	Ql	S
[149] Chevalier et al., 2017	56	5	12	Diadic	ASD	NAO	Care Home	Teens	Ql, Qn	A
[177] Clabaugh et al., 2019	41	14	17	Diadic	ASD	Kiwi	Home	Children	Qn	A

Citation	Per User Period (Days)	Per User # Sessions	Participants	HRI Type	Domain	Robot Used	Location	Age Group	Results	Operation
[93] Coninx et al., 2016	45	3	3	Diadic	Physical Health	NAO	Hospital	Children	Ql, Qn	N
[119] Coşar et al., 2020	70	daily	15	Diadic	Mental Health	Tiago	Home	Elderly	Ql, Qn	A
[134] Cruz-Sandoval et al., 2020	45	12	8	Group	Mental Health	Eva	Care Home	Elderly	Ql, Qn	A
[56] Davison et al., 2020	120	6	20	Diadic	Education	Zeno	School	Children	Qn	A
[152] de Graaf et al., 2015	30	22	6	Diadic	Physical Health	Karotz	Home	Elderly	Ql	A
[97] de Graaf et al., 2016	*unk.*	*unk.*	102	Mixed	General Purpose	Karotz	Home	Mixed	Qn	A
[96] de Graaf et al., 2017	*unk.*	*unk.*	102	Mixed	General Purpose	Karotz	Home	Mixed	Ql	A
[49] Donnermann et al., 2022	*unk.*	3	41	Diadic	Education	Pepper	School	Adults	Ql, Qn	A
[104] Edirisinghe & Jayasekara, 2018	7	*unk.*	4	Mixed	General Purpose	*unk.*	Lab	Mixed	Qn	A
[135] Fan et al., 2021	21	21	15	Diadic	Mental Health	NAO	Care Home	Elderly	Ql, Qn	N
[109] Fernaeus et al., 2010	180	*unk.*	6	Group	Entertainment	Pleo	Home	Children	Ql	A
[87] François et al., 2009	70	10	6	Triadic	ASD	Aibo	School	Children	Ql	S
[90] Gamborino et al., 2019	5	5	14	Diadic	Entertainment	ROBOHON	Lab	Children	Ql, Qn	A
[102] Gasteiger et al., 2021	6	free	6	Diadic	Mental Health	Bomy	Care Home	Elderly	Ql	A
[136] Gasteiger et al., 2021	84	15	10	Diadic	Mental Health	Bomy	Care Home	Elderly	Ql	A
[133] Giusti et al., 2006	30	8	5	Group	Mental Health	Paro	Care Home	Elderly	Ql	A
[146] Greczek et al., 2014	18	5	12	Diadic	ASD	NAO	School	Children	Ql, Qn	A
[155] Hebesberger et al., 2016	30	7	10	Group	Physical Health	SCITOS G5	Care Home	Elderly	Ql, Qn	A
[7] Hyun et al., 2010	14	10	111	Group	General Purpose	iRobiQ	Day Care	Children	Qn	A
[58] Irfan et al., 2020	126	35	1	Diadic	Physical Health	NAO	Facility	Adults	Ql, Qn	A
[4] J et al., 2013	14	4	33	Diadic	Physical Health	Bandit	Lab	Elderly	Qn	A
[129] Jacq et al., 2016	19	3.5	8	Diadic	Education	NAO	Facility	Children	Ql, Qn	A
[122] Jain et al., 2020	30	*unk.*	7	Diadic	ASD	Kiwi	Home	Children	Qn	A
[112] Janssen et al., 2011	14	3	20	Diadic	Education	NAO	School	Children	Qn	N
[65] Jeong et al., 2018	28	*unk.*	12	Mixed	General Purpose	Fribo	Home	Adults	Ql	A
[64] Jeong et al., 2020	12	7	35	Diadic	Mental Health	Jibo	Home	Adults	Ql, Qn	A
[8] Jeong et al., 2023	28	12	70	Diadic	Mental Health	Jibo	Home	Adults	Ql, Qn	A
[50] Jone et al., 2018	30	4	24	Diadic	Education	NAO	School	Children	Qn	A
[124] Kanda et al., 2004	14	9	228	Group	Education	Robovie	School	Children	Ql	A
[2] Kanda et al., 2007	32	*unk.*	37	Mixed	Entertainment	Robovie	School	Children	Ql, Qn	A
[52] Kidd et al., 2008	42	51	15	Diadic	Physical Health	Custom	Home	Mixed	Qn	A
[103] Klamer et al., 2011	10	40	3	Diadic	Mental Health	Nabaztag	Home	Adults	Ql	A
[92] Koay et al., 2007	35	8	12	Diadic	General Purpose	Custom	Lab	Adults	Qn	S
[161] Koay et al., 2016	35	10	9	Diadic	General Purpose	Sunflower	Lab	Adults	Ql, Qn	A
[55] Kory-Westlund et al., 2015	60	8	34	Diadic	General Purpose	Tega	School	Children	Qn	A
[128] Kory-Westlund et al., 2016	60	7	34	Diadic	Education	Tega	Day Care	Children	Ql	A
[85] Kozima et al., 2009	150	15	2	Mixed	General Purpose	Keepon	School	Children	Ql	N
[85] Kozima et al., 2009	*unk.*	20	27	Group	General Purpose	Keepon	School	Children	Ql	N
[85] Kozima et al., 2009	*unk.*	*unk.*	25	Diadic	General Purpose	Keepon	Lab	Children	Ql, Qn	N

Citation	Per User									
	Period (Days)	# Sessions	Participants	HRI Type	Domain	Robot Used	Location	Age Group	Results	Operation
[94] Lane et al., 2016	570	daily	23	Diadic	Mental Health	Paro	Facility	Elderly	Ql, Qn	A
[125] Leite et al., 2009	35	5	5	Diadic	Education	iCat	Other	Children	Ql, Qn	A
[131] Leite et al., 2014	35	5	16	Diadic	Education	iCat	School	Children	Ql, Qn	A
[54] Leite et al., 2015	21	3	40	Mixed	Education	Keepon	School	Children	Qn	A
[111] Leite et al., 2015	21	3	40	Mixed	Education	Keepon	School	Children	Qn	A
[118] Lemaignan et al., 2022	21	daily	30	Mixed	ASD	Pepper	School	Children	Ql, Qn	A
[162] Leyzberg et al., 2018	14	5	19	Diadic	Education	Keepon	School	Children	Qn	A
[141] Ligthart et al., 2022	60	5	46	Diadic	General Purpose	NAO	School	Children	Qn	A
[98] Ligthart et al., 2023	*unk.*	3	130	Diadic	Education	NAO	School	Children	Qn	A
[117] Luperto et al., 2022	14	daily	13	Diadic	Mental Health	Giraff-X	Home	Elderly	Ql, Qn	A
[157] McCallum & McOwan, 2015	42	6	10	Diadic	Entertainment	Mortimer	Lab	Adults	Qn	A
[120] Michaelis et al., 2018	14	daily	24	Diadic	Education	Maki	Home	Children	Ql, Qn	A
[167] Michaelis et al., 2023	28	27	6	Family	Education	Misty	Home	Children	Ql	A
[144] Michaud et al., 2007	49	22	2	Group	ASD	Puppet	Facility	Children	Ql	N
[158] Movellan et al., 2009	12	12	9	Diadic	Education	Custom	Day Care	Infants & Toddlers	Qn	A
[140] Nakanishi et al., 2022	9	18	40	Triadic	General Purpose	Sota	Day Care	Infants & Toddlers	Qt	N
[91] Nalin et al., 2012	35	3	13	Diadic	General Purpose	NAO	Hospital	Children	Ql, Qn	S
[116] Napoli et al., 2022	65	daily	7	Diadic	Mental Health	Sanbot Elf	Home	Elderly	Ql, Qn	A
[95] Nie et al., 2018	*unk.*	5	34	Diadic	ASD	NAO	Lab	Children	Qn	A
[110] Obayashi et al., 2022	180	daily	34	Group	Mental Health	Mon-chan	Care Home	Elderly	Ql, Qn	A
[6] Ostrowski et al., 2022	150	1559	28	Diadic	General Purpose	Jibo	Home	Elderly	Ql	A
[1] Paetzel et al., 2020	13	3	40	Diadic	General Purpose	Furhat	Lab	Adults	Qn	A
[615] Pakkar et al., 2019	30	*unk.*	8	Mixed	ASD	Kiwi	Home	Children	Ql	A
[66] Pelikan et al., 2020	11	*unk.*	20	Group	Entertainment	Cozmo	Home	Children	Ql	A
[115] Piasek et a., 2018	70	daily	10	Diadic	Mental Health	Tiago	Home	Elderly	Ql	A
[153] Polak & Levy-Tzedesk, 2020	41	15	4	Diadic	Physical Health	Pepper	Facility	Mixed	Ql	A
[148] Rakhymbayeva et al., 2021	21	10	11	Triadic	ASD	NAO	Facility	Children	Ql, Qn	A
[164] Ramachandran et al., 2016	14	4	29	Diadic	Education	NAO	School	Children	Qn	A
[127] Ramachandran et al., 2019	14	4	29	Diadic	Education	NAO	School	Children	Qn	A
[48] Ramachandran et al., 2019	21	5	28	Diadic	Education	NAO	School	Children	Ql, Qn	A
[59] Ramnauth et al., 2022	7	73	10	Diadic	ASD	Jibo	Home	Adults	Ql, Qn	A
[114] Rivoire et al., 2016	56	daily	10	Family	General Purpose	Pepper	Home	Adults	Ql, Qn	A
[107] Robins et al., 2005	180	9	4	Diadic	ASD	Robota	School	Children	Ql, Qn	N
[123] Rueben et al., 2021	42	5	6	Mixed	Service & Workplace	Custom	Other	Mixed	Ql	N
[121] Šabanović et al., 2014	28	daily	6	Diadic	Mental Health	Custom	Office	Adults	Ql, Qn	A
[143] Sabelli et al., 2011	105	23	55	Group	General Purpose	Robovie2	Care Home	Elderly	Ql	N
[137] Sahin et al., 2021	28	8	1	Diadic	Mental Health	NAO & Dash	Lab	Infants & Toddlers	Qn	N
[126] Salomons et al., 2022	14	14	14	Diadic	Mental Health	Keepon	Home	Adults	Qn	A
[99] Salter et al., 2004	*unk.*	5	8	Diadic	ASD	*unk.*	School	Children	Ql, Qn	A

	Per User									
Citation	**Period (Days)**	**# Sessions**	**Participants**	**HRI Type**	**Domain**	**Robot Used**	**Location**	**Age Group**	**Results**	**Operation**
[178] Sandygulova et al., 2022	21	7	34	Triadic	ASD	NAO	Facility	Children	Ql, Qn	S
[3] Scassellati et al., 2018	30	23	12	Triadic	ASD	Jibo	Home	Children	Ql, Qn	A
[84] Severinson et al., 2003	90	*unk.*	1	Diadic	General Purpose	Custom	Office	Adults	Ql	S
[159] Shi et al., 2022	30	5	4	Diadic	ASD	Kiwi	Home	Children	Qn	A
[150] Short et al., 2014	21	6	26	Diadic	Physical Health	DragonBot	School	Children	Qn	N
[106] Silvera-Tawil & Yates, 2018	150	*unk.*	45	Observer	ASD	NAO & Paro	School	Adults	Ql	A
[5] Silvera-Tawil et al., 2018	280	160	3	Mixed	ASD	NAO	School	Teens	Ql	A
[165] Singh et al., 2022	35	30	12	Triadic	Education	Cozmo	School	Children	Ql, Qn	N
[139] Spitale et al., 2023	28	4	26	Diadic	Mental Health	Misty & QT	Office	Adults	Ql, Qn	A
[89] Stubbs et al., 2005	105	*unk.*	11	Observer	Education	PER Rover	Other	Adults	Qn	A
[142] Sung et al., 2010	75	daily	48	Family	General Purpose	Roomba	Home	Adults	Ql, Qn	A
[151] Süssenbach et al., 2014	18	18	16	Diadic	Physical Health	NAO	Lab	Adults	Ql, Qn	A
[108] Tanaka et al., 2007	150	45	11	Group	Entertainment	QRIO	Day Care	Infants & Toddlers	Ql, Qn	S
[156] Taylor et al., 2021	10	20	9	Observer	Entertainment	Custom	Day Care	Adults	Ql	S
[86] Tolksdorf et al., 2020	14	4	29	Observer	Education	NAO	Lab	Mixed	Ql	N
[57] Trinh et al., 2020	3	5	20	Diadic	Physical Health	Patterns	Hospital	Elderly	Ql, Qn	A
[88] Vishwanath et al., 2019	30	*unk.*	12	Observer	Service & Workplace	Nadine	Office	Adults	Ql	A
[160] Vogt et al., 2019	21	7	108	Diadic	Education	NAO	School	Children	Qn	S
[132] Wada et al., 2006	52	*unk.*	12	Group	Mental Health	Paro	Care Home	Elderly	Ql, Qn	A
[60] Wada et al., 2007	52	*unk.*	12	Group	Mental Health	Paro	Care Home	Elderly	Ql, Qn	A
[101] Wada et al., 2013	7	*unk.*	80	Diadic	Mental Health	Paro	Care Home	Elderly	Ql	A
[138] Wada et al., 2014	30	*unk.*	64	Diadic	Mental Health	Paro	Care Home	Elderly	Ql	A
[42] Weiss et al., 2021	210	*unk.*	8	Family	General Purpose	Vector	Home	Mixed	Ql	A
[145] Zhanatykyzy et al., 2023	21	6	34	Triadic	ASD	NAO	Facility	Children	Qn	N
[166] Zhang et al., 2023	21	8	31	Diadic	Education	Jibo	Home	Children	Qn	A
[44] Zhao & McEwen, 2022	180	75	27	Diadic	Education	Luka	Home	Children	Ql, Qn	S

Appendix B: Robots for Autism Therapy Review Corpus and Summary Table

The table below provides a high-level summary of the data and key study characteristics extracted from the review presented in Chapter 3. To conserve space and improve readability, a set of standardized abbreviations is used throughout the table.

The "Venue" column indicates the domain of the publication, with **T** referring to technical venues (e.g., robotics, AI, HRI), **C** to clinical venues (e.g., medical or psychological journals), and **I** to interdisciplinary or other venues (e.g., education, rehabilitation).

Study design types are abbreviated as follows: **Case or SS** for case studies or single-subject designs, **Exp.|RCT** for randomized controlled trials and **Exp.|Non-RCT** for experimental design without a randomized control, **Obs.|CS** for observational cross-sectional studies, **Pilot** for pilot or feasibility studies, **Method** for methodological or validation work, and **System** for system/prototype descriptions. **Other|*unk.*** refers to study designs that were either unconventional or not clearly reported by the authors.

The "Targeted Skills" column reflects the focus of the intervention. **Cog.** refers to cognitive skills such as reasoning and problem-solving, **JA** to joint attention, **Eng.** to general social engagement, **Im.** to imitation, **Gaze** to visual attention or gaze-following, **Com.|NV** to non-verbal communication, and **Com.|V** to verbal communication. **Emo.** captures emotion recognition or expression, **TT** denotes turn-taking, **Motor** refers to motor skills or coordination, **Stereotypy** reflects efforts to reduce repetitive behaviors, and **Sensory** relates to sensory engagement.

Participant age groups are coded as **Early** for early childhood (0–5 years), **Middle** for middle childhood (6–12 years), **Teen** for adolescence (13–17 years), **Adult** for participants 18 years or older, and **Mixed** for studies spanning multiple age ranges.

The robot column lists either the robot platform (e.g., NAO, KASPAR, Keepon) or **Custom** for bespoke robots. **Other|G** refers to general-purpose platforms. Autonomy is marked as **A** for fully autonomous operation, **S** for partially or semi-autonomous, and **N** for non-autonomous control.

The robot's role is denoted as **Peer** (social companion), **Trainer** (instructor, therapist, or coach), **Mediator** (facilitator of social interaction), or **Other|G** for generalized roles.

Lastly, the "Interaction Structure" column describes the social configuration of the intervention: **Dyadic** indicates one-on-one interaction between the child and the robot; **Triadic|Ther.** involves a child–robot–therapist configuration; **Triadic|Peer** includes a peer who mediates or participates in the interaction; **Triadic|Parent** denotes caregiver involvement; and **Triadic|Other** refers to other triadic arrangements not captured by the previous categories. Group-based or nonstandard configurations are listed as **Other**.

The label ***unk.*** is used throughout to indicate missing or unreported data.

It is important to note that many labels in this summary table have been grouped or standardized for clarity and space efficiency. As a result, they may not fully capture the nuance or specificity reported in individual studies. For detailed descriptions and contextual analysis, please refer to the corresponding sections in Chapter 3.

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[293]	Abu-Amara et al., 2024	T	Exp.|Non-RCT	Other|*unk.*	Clinic	Middle	NAO	S	Trainer	Dyadic
[616]	Ackovska et al., 2023	T	Exp.|Non-RCT	Daily Living Skills, Gaze, Im., TT, Com.|V	Clinic	Mixed	KASPAR	S	Trainer	Dyadic
[617]	Alarcon et al., 2021	T	Qualitative	Cog.	School	Middle	NAO	A	Trainer	Dyadic
[618]	Albo-Canals et al., 2013	T	Exp.|Non-RCT	Eng., JA, Eng.	School	Middle	LEGO	S	Peer	Other
[619]	Ali et al., 2019	T	Exp.|Non-RCT	Cog., Im., JA	Clinic	Mixed	NAO	A	Trainer	Dyadic
[620]	Ali et al., 2020	T	Exp.|Non-RCT	Cog., JA	Clinic	Mixed	NAO	S	Peer	Other
[621]	Ali et al., 2020	T	Exp.|Non-RCT	Gaze, JA	Clinic	Middle	NAO	S	Trainer	Dyadic
[622]	Ali et al., 2022	C	Exp.|Non-RCT	Cog., Im., JA	Clinic	Early	NAO	S	Trainer	Dyadic
[623]	Al-Nafjan et al., 2023	C	Exp.|Non-RCT	Emo., Eng.	Clinic	Middle	*unk.*	S	Trainer	Dyadic
[624]	Alnajjar et al., 2021	T	Exp.|Non-RCT	Cog., JA, Eng.	Clinic	Middle	NAO	S	Trainer	Dyadic
[625]	Amirabdollahian et al., 2011	T	Obs.|CS	Sensory	School	Early	KASPAR	S	Peer	Dyadic
[626]	Amirova et al., 2023	C	Exp.|Non-RCT	Gaze, Eng.	Clinic	Middle	NAO	S	Peer	Triadic|Ther.
[627]	Anamaria et al., 2013	T	Exp.|Non-RCT	Emo.	Clinic	Early	Probo	S	Peer	Dyadic
[628]	Andreae et al., 2014	I	Exp.|Non-RCT	Eng., Motor, Com.|V	Home	Middle	Auti	S	Peer	Dyadic
[629]	Annunziata et al., 2024	C	Exp.|Non-RCT	Im., Motor, Com.|NV	Clinic	Early	NAO	S	Peer	Triadic|Ther.
[630]	Anzalone et al., 2014	C	Exp.|Non-RCT	Gaze, Eng., JA, Eng.	Lab	Mixed	NAO	S	Trainer	Dyadic
[631]	Attawibulkul et al., 2019	I	Exp.|Non-RCT	Other|*unk.*	School	Middle	BLISS	S	Peer	Triadic|Ther.
[632]	Axelsson et al., 2019	T	Exp.|Non-RCT	Cog., Eng., Im., Com.|V	Clinic	Mixed	InMoov	S	Trainer	Dyadic
[633]	Aziz et al., 2015	T	Exp.|Non-RCT	Other|*unk.*	Lab	Early	NAO	S	Peer	Dyadic
[634]	Baraka et al., 2020	T	Case or SS	Other|*unk.*	Clinic	Early	NAO	A	Trainer	Dyadic
[635]	Baraka et al., 2022	T	Exp.|Non-RCT	Cog., JA	Clinic	Early	NAO	N	Trainer	Triadic|Ther.
[636]	Barakova et al., 2015	T	Case or SS	Eng., TT	School	Middle	NAO	S	Peer	Other
[637]	Barnes et al., 2021	T	Exp.|Non-RCT	Cog., Im., Eng., Motor	Lab	Mixed	NAO	S	Other|G	Dyadic
[638]	Begum et al., 2015	T	Case or SS	Eng.	Facility	Teen	NAO	S	Trainer	Triadic|Parent
[639]	Bekele et al., 2011	I	System	Cog., JA	Lab	Early	NAO	S	Trainer	Dyadic
[640]	Bekele et al., 2011	T	Exp.|Non-RCT	Cog., JA	Lab	Early	NAO	S	Trainer	Triadic|Peer
[641]	Bekele et al., 2013	C	Case or SS	Gaze, JA	Clinic	Early	NAO	S	Trainer	Triadic|Peer
[642]	Berk-Smeekens et al., 2020	I	Exp.|Non-RCT	Eng.	Clinic	Early	NAO	S	Trainer	Triadic|Parent
[333]	Berk-Smeekens et al., 2022	C	Exp.|RCT	Eng.	Clinic	Early	NAO	S	Trainer	Triadic|Ther.
[643]	Bharatharaj et al., 2016	T	Exp.|Non-RCT	Cog., Emo., Eng.	School	Middle	KiliRo	S	Peer	Triadic|Peer
[644]	Bharatharaj et al., 2017	T	Exp.|Non-RCT	Cog., Im., JA	Other|*unk.*	Middle	KiliRo	S	Peer	Dyadic

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[645]	Bharatharaj et al., 2017	T	Pilot or Feasibility	Cog., Eng.	Clinic	Middle	KiliRo	S	Peer	Dyadic
[646]	Bharatharaj et al., 2017	I	Exp.|Non-RCT	Emo., Eng.	School	Middle	KiliRo	S	Peer	Dyadic
[319]	Billard et al., 2006	T	Longitudinal	Cog., Im., JA, TT	Lab	Mixed	Robota	A	Peer	Triadic|Other
[341]	Billing et al., 2020	I	Other|*unk.*	Cog., Im., JA, TT	Clinic	Early	NAO	A	Trainer	Triadic|Ther.
[647]	Bird et al., 2007	I	Exp.|Non-RCT	Im.	Lab	Adult	*unk.*	A	Other|G	Dyadic
[648]	Boccanfuso et al., 2016	T	Exp.|Non-RCT	Cog., Im., JA, TT, Com.|V	Lab	Early	CHARLIE	S	Peer	Triadic|Other
[649]	Brienza et al., 2023	T	Case or SS	JA	Other|*unk.*	Early	NAO	S	Trainer	Dyadic
[650]	Bugnariu et al., 2013	C	Exp.|Non-RCT	Im., JA, Motor	Clinic	Middle	NAO	N	Trainer	Triadic|Other
[651]	Cai et al., 2019	T	Method	Cog., Im., JA, TT	Lab	Early	NAO	S	Trainer	Triadic|Peer
[652]	Cao et al., 2019	T	Other|*unk.*	Cog., Eng., Im., Eng.	Clinic	Early	NAO	S	Trainer	Dyadic
[340]	Cao et al., 2019	T	Exp.|Non-RCT	Cog., Im., JA, TT	Clinic	Early	NAO	S	Trainer	Triadic|Ther.
[653]	Cao et al., 2022	T	Exp.|Non-RCT	Cog., Im.	Clinic	Early	NAO	S	Other|G	Dyadic
[654]	Casas-Bocanegra et al., 2020	T	Case or SS	Cog., JA, Motor	Clinic	Middle	Custom	S	Peer	Dyadic
[655]	Cervera et al., 2018	T	Exp.|RCT	Com.|NV, Com.|V	Clinic	Early	NAO	N	Trainer	Triadic|Parent
[656]	Chen et al., 2021	T	System	Cog., Emo.	School	Mixed	Custom	A	Trainer	Dyadic
[657]	Chevalier et al., 2016	T	Exp.|Non-RCT	Gaze, JA	Clinic	Middle	NAO	N	Trainer	Dyadic
[658]	Chevalier et al., 2017	T	Exp.|Non-RCT	Im., Motor	Other|*unk.*	Middle	NAO	S	Trainer	Dyadic
[659]	Chevalier et al., 2022	T	Exp.|Non-RCT	Cog., JA	Clinic	Early	Cozmo	S	Trainer	Dyadic
[660]	Chung et al., 2019	C	Exp.|Non-RCT	Gaze, Com.|V	School	Middle	NAO	S	Trainer	Triadic|Peer
[296]	Chung et al., 2021	C	Exp.|Non-RCT	Cog., JA, Com.|V	School	Middle	NAO	S	Trainer	Triadic|Ther.
[295]	Clabaugh et al., 2019	T	Exp.|Non-RCT	Other|*unk.*	Home	Early	Custom	A	Trainer	Dyadic
[283]	Conn et al., 2008	T	Exp.|Non-RCT	Emo., Eng.	Lab	Teen	Custom	S	Other|G	Dyadic
[661]	Conti et al., 2015	T	Case or SS	Cog., Im., Eng.	Clinic	Middle	NAO	A	Trainer	Triadic|Peer
[662]	Conti et al., 2019	T	Exp.|Non-RCT	Emo.	Other|*unk.*	Middle	NAO	S	Trainer	Dyadic
[663]	Coşkun et al., 2022	T	Exp.|Non-RCT	Emo.	Other|*unk.*	Middle	KASPAR	S	Peer	Triadic|Ther.
[664]	Costa et al., 2009	T	Exp.|Non-RCT	Cog.	School	Teen	KASPAR	S	Other|G	Dyadic
[265]	Costa et al., 2010	T	Exp.|Non-RCT	Cog., JA, Motor, Sensory, TT	School	Teen	LEGO	S	Peer	Triadic|Ther.
[665]	Costescu et al., 2016	C	Other|*unk.*	Other|*unk.*	Other|*unk.*	Middle	Keepon	S	Other|G	Dyadic
[666]	Costescu et al., 2017	C	Exp.|Non-RCT	Other|*unk.*	Facility	Middle	Keepon	S	Trainer	Dyadic
[208]	Dautenhahn et al., 2002	T	Exp.|Non-RCT	Gaze	School	Middle	Labo-1	A	Peer	Dyadic
[667]	David et al., 2018	T	Case or SS	Cog., JA	Other|*unk.*	Early	NAO	S	Trainer	Triadic|Peer
[668]	David et al., 2020	C	Exp.|Non-RCT	TT	Other|*unk.*	Early	NAO	S	Other|G	Triadic|Peer

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[669]	Dehkordi et al., 2015	T	Obs.|CS	Other|*unk.*	Facility	Early	Custom	S	Other|G	Dyadic
[670]	Desideri et al., 2017	I	Exp.|Non-RCT	Im., Eng.	Clinic	Early	NAO	S	Trainer	Dyadic
[671]	Desideri et al., 2018	C	Case or SS	Im., Motor, Com.|V	Clinic	Middle	NAO	N	Other|G	Dyadic
[672]	Dimitrova et al., 2012	T	Exp.|Non-RCT	Eng., Im.	School	Mixed	AdMoVeo	S	Peer	Triadic|Peer
[673]	Does et al., 2023	T	Other|*unk.*	Cog.	School	Mixed	*unk.*	S	Trainer	Dyadic
[266]	Duquette et al., 2008	T	Case or SS	Cog., JA, Eng., Com.|NV	Clinic	Mixed	Custom	S	Peer	Triadic|Ther.
[674]	Ercolano et al., 2024	I	Exp.|Non-RCT	Im., Motor, Com.|NV	Clinic	Mixed	NAO	S	Trainer	Dyadic
[675]	Esteban et al., 2017	T	Exp.|Non-RCT	Cog., Im., JA, TT	Clinic	Mixed	NAO, Probo, iRobiQ, CARO	S	Trainer	Triadic|Other
[676]	Fachantidis et al., 2020	C	Other|*unk.*	Gaze, Eng.	School	Middle	Custom	S	Peer	Other
[677]	Fachantidis et al., 2020	C	Case or SS	Gaze, Com.|V	Other|*unk.*	Middle	Daisy	S	Trainer	Dyadic
[678]	Febtriko et al., 2018	T	Exp.|Non-RCT	Eng.	School	Mixed	Custom	S	Other|G	Dyadic
[679]	Feil-Seifer et al., 2008	T	Case or SS	Cog., Eng., TT	Lab	Mixed	Custom	A	Peer	Dyadic
[254]	Feil-Seifer et al., 2009	T	Exp.|Non-RCT	Gaze, JA, Com.|V	Clinic	Mixed	Custom	S	Peer	Triadic|Ther.
[680]	Feil-Seifer et al., 2011	T	Exp.|Non-RCT	Eng.	Lab	Mixed	Custom	A	Peer	Triadic|Other
[681]	Feil-Seifer et al., 2012	T	Exp.|Non-RCT	Other|*unk.*	Lab	Mixed	Custom	A	Peer	Triadic|Other
[682]	Feng et al., 2017	T	Other|*unk.*	Im., Eng.	Lab	Early	NAO	A	Trainer	Dyadic
[337]	Feng et al., 2022	T	Exp.|Non-RCT	Emo., Motor, TT	Lab	Mixed	NAO	A	Trainer	Dyadic
[683]	Fournier et al., 2024	I	Exp.|Non-RCT	Im.	Clinic	Early	Pepper	A	Trainer	Triadic|Ther.
[684]	François et al., 2009	I	Longitudinal	Cog., Emo., JA	School	Mixed	Custom	A	Peer	Triadic|Other
[685]	Fuentes-Alvarez et al., 2023	T	Case or SS	Other|*unk.*	Lab	Teen	AR4A	A	Trainer	Dyadic
[273]	Gaitán-Padilla et al., 2022	T	Case or SS	Cog., Emo., Im.	Clinic	Middle	CASTOR	N	Peer	Dyadic
[686]	Galán-Mena et al., 2016	T	System	Eng.	Clinic	Mixed	Custom	S	Trainer	Other
[687]	Ghiglino et al., 2021	C	Exp.|RCT	Eng.	Clinic	Early	Cozmo	A	Peer	Dyadic
[236]	Giannopulu et al., 2010	C	Case or SS	Gaze, Motor, Com.|NV	Clinic	Middle	Custom	A	Peer	Other
[294]	Giannopulu et al., 2012	I	Case or SS	Emo.	Clinic	Early	Custom	A	Peer	Other
[688]	Giannopulu et al., 2013	I	Case or SS	Gaze, Motor, Com.|NV, Com.|V	Clinic	Middle	Custom	A	Peer	Triadic|Ther.
[689]	Greczek et al., 2014	T	Exp.|Non-RCT	Im.	School	Middle	NAO	A	Trainer	Dyadic
[690]	Hirokawa et al., 2016	T	Method	Emo., Eng.	Clinic	Middle	NAO	S	Peer	Dyadic
[691]	Holeva et al., 2024	C	Exp.|RCT	Stereotypy, Emo., Eng.	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[274]	Huijnen et al., 2021	T	Exp.|Non-RCT	Cog., Im., Motor, Com.|NV, Com.|V	School	Mixed	KASPAR	S	Trainer	Triadic|Other
[237]	Huskens et al., 2012	C	Exp.|Non-RCT	Other|*unk.*	School	Middle	NAO	A	Trainer	Dyadic
[277]	Huskens et al., 2015	C	Case or SS	TT, Com.|V	Clinic	Mixed	NAO	A	Trainer	Triadic|Peer
[692]	Ilijoski et al., 2022	T	Other|*unk.*	Cog., Im., JA, Com.|NV	Clinic	Mixed	KASPAR	A	Trainer	Triadic|Ther.
[693]	Ishak et al., 2019	T	Other|*unk.*	Im.	Clinic	Mixed	Rero	A	Trainer	Triadic|Ther.
[694]	Ismail et al., 2012	I	Exp.|Non-RCT	Gaze, Com.|V	Clinic	Mixed	NAO	A	Peer	Dyadic
[695]	Ismail et al., 2012	I	Exp.|Non-RCT	Motor	Clinic	Mixed	NAO	A	Trainer	Dyadic
[696]	Ivani et al., 2022	I	System	Im., Motor, Com.|NV, TT	Clinic	Early	NAO	S	Trainer	Triadic|Ther.
[344]	Jain et al., 2020	T	Longitudinal	Cog., Emo., Eng., Com.|V	Home	Mixed	Kiwi	A	Peer	Dyadic
[697]	Javed et al., 2018	T	Pilot or Feasibility	Cog., Eng.	Clinic	Mixed	NAO	A	Trainer	Dyadic
[357]	Javed et al., 2019	T	Exp.|Non-RCT	Im., Eng., TT	Clinic	Mixed	Romo	S	Trainer	Dyadic
[698]	Javed et al., 2020	T	Other|*unk.*	Emo., Eng., Motor, Com.|NV	Clinic	Mixed	Romo	S	Trainer	Dyadic
[699]	Jordan et al., 2013	C	Exp.|Non-RCT	Stereotypy, TT	School	Teen	iROBi-Q	S	Trainer	Triadic|Peer
[700]	Kaboski et al., 2015	C	Other|*unk.*	Cog., JA, Eng., TT	School	Mixed	LEGO	A	Peer	Other
[701]	Karakosta et al., 2019	T	Other|*unk.*	Gaze, Eng., Im., Eng., Motor, Com.|NV, Com.|V	School	Mixed	KASPAR	S	Peer	Triadic|Ther.
[702]	Karim et al., 2023	T	Other|*unk.*	Com.|NV	Clinic	Mixed	NAO	S	Trainer	Dyadic
[62]	Kim et al., 2012	C	Exp.|Non-RCT	Cog., Emo., Eng., TT, Com.|V	Clinic	Mixed	Pleo	S	Peer	Triadic|Peer
[238]	Kim et al., 2012	T	Exp.|Non-RCT	Emo., Eng.	Clinic	Mixed	Pleo	S	Trainer	Dyadic
[703]	Kim et al., 2014	T	Case or SS	Eng.	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.
[704]	Kim et al., 2021	T	Case or SS	Eng.	Clinic	Mixed	NAO	S	Other|G	Triadic|Ther.
[705]	Kim et al., 2022	T	Case or SS	Eng.	Clinic	Mixed	NAO	S	Other|G	Triadic|Parent
[706]	Kim et al., 2024	T	Case or SS	Eng., Com.|V	School	Teen	Sota	S	Other|G	Triadic|Peer
[707]	Koch et al., 2017	T	Exp.|Non-RCT	Emo.	Clinic	Mixed	SAM	A	Trainer	Dyadic
[708]	Konishi et al., 2024	C	Exp.|Non-RCT	Emo., Eng	Clinic	Adult	Android ST	S	Other|G	Dyadic
[709]	Korneder et al., 2022	T	Case or SS	Com.|V	Clinic	Mixed	NAO	S	Trainer	Dyadic
[710]	Korte et al., 2020	C	Exp.|RCT	Other|*unk.*	School	Early	NAO	S	Other|G	Triadic|Ther.

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[711]	Kostrubiec et al., 2020	T	Other|*unk.*	TT	School	Mixed	Custom	A	Trainer	Triadic|Peer
[302]	Kostrubiec et al., 2024	T	Exp.|Non-RCT	Eng.	Clinic	Mixed	Nao	S	Trainer	Triadic|Ther.
[233]	Kozima et al., 2007	C	Case or SS	Emo., Gaze, JA, Eng.	Clinic	Early	Keepon	A	Other|G	Triadic|Parent
[270]	Kozima et al., 2009	T	Case or SS	Emo., Gaze, Im., JA	School	Mixed	Keepon	A	Other|G	Triadic|Ther.
[288]	Kumazaki et al., 2017	C	Exp.|Non-RCT	Other|*unk.*	Clinic	Adult	Actroid-F	S	Other|G	Dyadic
[712]	Kumazaki et al., 2017	I	Exp.|Non-RCT	Other|*unk.*	Clinic	Teen	ACTROID-F, M3-Synchy	S	Other|G	Dyadic
[713]	Kumazaki et al., 2018	C	Exp.|Non-RCT	Cog., JA	Clinic	Mixed	CommU	S	Trainer	Dyadic
[300]	Kumazaki et al., 2019	C	Exp.|Non-RCT	Cog., Eng., JA	Clinic	Mixed	CommU	A	Trainer	Triadic|Peer
[714]	Kumazaki et al., 2019	C	Exp.|Non-RCT	Cog., Eng.	Clinic	Early	Actroid-F, CommU	S	Other|G	Other
[289]	Kumazaki et al., 2019	C	Exp.|Non-RCT	Motor, Com.|NV, Com.|V	Clinic	Adult	Actroid-F	S	Other|G	Dyadic
[286]	Kumazaki et al., 2019	C	Exp.|Non-RCT	Cog., Eng.	Clinic	Adult	Actroid-F	S	Other|G	Dyadic
[715]	Kumazaki et al., 2021	C	Exp.|Non-RCT	Other|*unk.*	Clinic	Adult	CommU	S	Peer	Triadic|Peer
[716]	Kumazaki et al., 2022	C	Exp.|Non-RCT	Sensory	Clinic	Teen	A-Lab ST	S	Other|G	Dyadic
[717]	Kwon et al., 2015	T	Case or SS	Gaze, Im., Motor, Com.|V	Clinic	Mixed	Custom	S	Other|G	Dyadic
[718]	Lakatos et al., 2021	T	Exp.|Non-RCT	Other|*unk.*	School	Middle	KASPAR	S	Trainer	Dyadic
[719]	Lecciso et al., 2021	C	Other|*unk.*	Com.|NV	Clinic	Middle	Zeno	S	Trainer	Dyadic
[720]	Lee et al., 2012	T	Exp.|Non-RCT	Gaze, Com.|NV, Com.|V	Clinic	Middle	Ifbot	S	Trainer	Dyadic
[282]	Lee et al., 2012	T	Exp.|Non-RCT	Motor	Clinic	Mixed	Custom	A	Trainer	Dyadic
[721]	Lee et al., 2013	T	Exp.|Non-RCT	Cog., Gaze	Clinic	Adult	Custom	A	Other|G	Dyadic
[722]	Lee et al., 2013	T	Exp.|Non-RCT	Motor	Clinic	Mixed	Custom	S	Other|G	Dyadic
[723]	Lee et al., 2014	T	Exp.|Non-RCT	Emo., Eng., Motor	Clinic	Mixed	Ifbot	S	Trainer	Dyadic
[336]	Lee et al., 2021	I	Exp.|Non-RCT	Other|*unk.*	Clinic	Mixed	NAO	S	Peer	Triadic|Ther.
[284]	Lemaignan et al., 2022	T	Longitudinal	Eng.	School	Teen	Pepper	S	Peer	Triadic|Peer
[724]	Lin et al., 2022	T	Case or SS	Other|*unk.*	Clinic	Early	Custom	S	Peer	Triadic|Ther.
[725]	Liu et al., 2007	T	Exp.|Non-RCT	Eng.	Clinic	Teen	Custom	A	Other|G	Dyadic
[726]	Liu et al., 2008	T	Exp.|Non-RCT	Emo., Eng.	Clinic	Teen	CRS Catalyst-5	A	Trainer	Dyadic
[727]	Liu et al., 2016	T	Exp.|Non-RCT	Im., Motor	Clinic	Mixed	NAO	A	Peer	Triadic|Peer
[728]	Lorenzo et al., 2024	T	Exp.|Non-RCT	Emo., Eng.	School	Mixed	NAO	S	Peer	Triadic|Ther.

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[729]	Louie et al., 2021	T	Case or SS	Other|*unk.*	Clinic	Early	NAO	N	Trainer	Dyadic
[730]	Lund et al., 2009	T	Case or SS	Other|*unk.*	Clinic	Mixed	Custom	A	Other|G	Triadic|Ther.
[731]	Lytridis et al., 2022	T	Case or SS	Gaze, Eng.	Clinic	Mixed	NAO	A	Other|G	Triadic|Ther.
[732]	Malik et al., 2013	I	Pilot or Feasibility	Cog., Im.	Clinic	Mixed	NAO	S	Trainer	Dyadic
[733]	Manner et al., 2015	T	Pilot or Feasibility	Cog., Eng., Im.	Clinic	Early	NAO	S	Trainer	Triadic|Parent
[276]	Marathaki et al., 2022	C	Exp.|Non-RCT	Gaze, Im.	School	Mixed	NAO	S	Trainer	Triadic|Ther.
[335]	Marino et al., 2020	C	Exp.|RCT	Other|*unk.*	Clinic	Mixed	NAO	S	Trainer	Other
[734]	Martínez et al., 2022	T	Case or SS	Cog., Motor, Com.|NV	School	Mixed	Custom	S	Peer	Triadic|Peer
[735]	Mavadati et al., 2014	T	Exp.|Non-RCT	JA, Com.|V	Clinic	Teen	NAO	S	Other|G	Dyadic
[736]	Mavadati et al., 2016	T	Case or SS	Cog., JA, Com.|NV, Com.|V	Clinic	Mixed	NAO	S	Trainer	Triadic|Peer
[737]	Mayadunne et al., 2020	T	Exp.|Non-RCT	Motor	Clinic	Mixed	Custom	S	Peer	Triadic|Ther.
[738]	Mazzei et al., 2010	T	Case or SS	Emo., Emo., Im., Eng.	Lab	Mixed	FACE	S	Trainer	Triadic|Ther.
[272]	Mazzei et al., 2011	T	Exp.|Non-RCT	Cog., Emo., Im.	Lab	Adult	FACE	S	Trainer	Triadic|Ther.
[739]	Mazzei et al., 2012	T	Exp.|Non-RCT	Cog., Emo.	Lab	Mixed	FACE	A	Trainer	Triadic|Ther.
[740]	Mehmood et al., 2021	T	Exp.|Non-RCT	Cog.	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.
[741]	Mehralizadeh et al., 2023	T	Exp.|Non-RCT	Stereotypy, Eng., Sensory	Clinic	Mixed	Custom	S	Trainer	Triadic|Ther.
[281]	Melo et al., 2019	I	Longitudinal	Cog., Eng., TT	Clinic	Mixed	Astro	A	Trainer	Triadic|Ther.
[742]	Meltzoff et al., 2010	I	Exp.|RCT	Gaze, JA	Lab	Early	HOAP-2	A	Peer	Dyadic
[743]	Mengoni et al., 2017	I	Exp.|RCT	Cog., Eng., Im., JA, TT	Clinic	Mixed	KASPAR	*unk.*	Trainer	Triadic|Ther.
[744]	Moorthy et al., 2016	T	Exp.|Non-RCT	Im., Motor	Clinic	Mixed	LEGO	A	Trainer	Dyadic
[256]	Nakadoi et al., 2017	T	Case or SS	Emo.	Clinic	Mixed	PARO	A	Peer	Triadic|Peer
[745]	Niderla et al., 2021	T	Pilot or Feasibility	Emo., Eng.	Clinic	Mixed	Custom	S	Peer	Triadic|Ther.
[746]	Nie et al., 2018	T	Exp.|Non-RCT	Cog., JA	Clinic	Early	NAO	S	Trainer	Dyadic
[747]	Nie et al., 2024	T	Exp.|RCT	JA, Eng.	Clinic	Early	NORRIS	A	Trainer	Triadic|Ther.
[748]	Nunez et al., 2015	T	Case or SS	Eng., TT	Clinic	Mixed	Custom	S	Peer	Dyadic
[749]	Nuovo et al., 2018	T	Other|*unk.*	Cog., Gaze, Im.	Clinic	Mixed	NAO	S	Trainer	Dyadic
[750]	Oliver et al., 2019	T	Obs.|CS	Cog., Im., JA, TT	Clinic	Mixed	Cozmo	*unk.*	Trainer	Triadic|Parent
[751]	Otterdijk et al., 2020	T	Other|*unk.*	Cog., Eng.	Other|*unk.*	Early	NAO	S	Peer	Triadic|Parent
[752]	Paengkumhag et al., 2023	I	Other|*unk.*	Cog., Eng.	Clinic	Mixed	BLISS	A	Trainer	Triadic|Peer
[753]	Pakkar et al., 2019	T	Case or SS	Cog., Emo., JA	Home	Mixed	SPRITE	A	Trainer	Dyadic
[754]	Palestra et al., 2017	I	Case or SS	Cog., Gaze	Clinic	Middle	NAO	A	Trainer	Dyadic

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[755]	Palestra et al., 2017	T	Exp.|Non-RCT	Cog., JA, Com.|V	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.
[756]	Panceri et al., 2021	T	Exp.|Non-RCT	Stereotypy, Cog., JA, Motor, Com.|V	Clinic	Mixed	Custom	A	Trainer	Triadic|Ther.
[757]	Papazoglou et al., 2021	T	Exp.|Non-RCT	Eng.	School	Mixed	LEGO	*unk.*	Peer	Other
[758]	Peca et al., 2014	T	Exp.|Non-RCT	Cog.	School	Mixed	Keepon, Nao, Probo, Pleo, KASPAR, Romibo	A	Other|G	Dyadic
[759]	Pérez-Vázquez et al., 2023	T	Case or SS	Com.|V	School	Mixed	Bee-Bot	*unk.*	Trainer	Dyadic
[760]	Peribañez et al., 2023	T	Exp.|Non-RCT	Cog.	Clinic	Mixed	Ozobot	*unk.*	Trainer	Dyadic
[761]	Petric et al., 2017	T	Exp.|Non-RCT	Cog., Im., JA, Com.|NV	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.
[762]	Pierno et al., 2008	I	Exp.|Non-RCT	Im., Motor	Clinic	Mixed	Custom	S	Trainer	Dyadic
[763]	Pinto-Bernal et al., 2022	T	Other|*unk.*	Gaze	Clinic	Mixed	CASTOR	S	Trainer	Triadic|Ther.
[764]	Pioggia et al., 2005	T	Case or SS	Cog., Emo., Im.	Clinic	Middle	FACE	S	Trainer	Dyadic
[255]	Pioggia et al., 2007	T	Exp.|Non-RCT	Gaze, Im.	Clinic	Mixed	FACE	A	Trainer	Dyadic
[234]	Pioggia et al., 2008	C	Exp.|Non-RCT	Cog., Im.	Clinic	Mixed	FACE	S	Trainer	Dyadic
[765]	Pioggia et al., 2022	T	Exp.|Non-RCT	Other|*unk.*	Lab	Mixed	QT	S	Trainer	Other
[280]	Pliasa et al., 2019	T	Exp.|Non-RCT	TT, Com.|V	School	Early	Daisy	S	Mediator	Triadic|Peer
[766]	Pop et al., 2013	C	Case or SS	Cog., Im., Motor	Clinic	Early	Robonova-1	A	Trainer	Dyadic
[767]	Pop et al., 2013	C	Exp.|Non-RCT	Gaze, Eng., Com.|V	Clinic	Mixed	Probo	A	Trainer	Triadic|Ther.
[768]	Pop et al., 2014	I	Exp.|Non-RCT	Gaze, Eng., Com.|V	Clinic	Mixed	Probo	A	Other|G	Triadic|Ther.
[769]	Pour et al., 2018	T	Exp.|Non-RCT	Cog., Emo., Im., JA, Com.|NV	Clinic	Early	R50-Alice	S	Peer	Dyadic
[770]	Pradel et al., 2010	T	Pilot or Feasibility	Cog., Im., Eng., Sensory	Clinic	Mixed	Custom	A	Mediator	Triadic|Ther.
[771]	Puyon et al., 2013	T	Exp.|Non-RCT	Cog., Gaze, Com.|V	Clinic	Mixed	Custom	A	Mediator	Dyadic
[772]	Qidwai et al., 2020	T	Exp.|Non-RCT	Cog.	School	Middle	NAO	S	Peer	Triadic|Other
[773]	Rakhymbayeva et al., 2021	T	Exp.|Non-RCT	Gaze, Im., Eng.	Clinic	Mixed	NAO	A	Mediator	Triadic|Ther.
[774]	Ramírez-Duque et al., 2018	T	Obs.|CS	Gaze, JA	Clinic	Mixed	ONO	S	Trainer	Triadic|Ther.
[775]	Ramírez-Duque et al., 2019	T	Obs.|CS	Cog., Im., JA	Clinic	Mixed	ONO	S	Mediator	Triadic|Ther.
[776]	Ramírez-Duque et al., 2020	T	Exp.|Non-RCT	Gaze, JA	Clinic	Mixed	ONO	S	Trainer	Triadic|Ther.
[29]	Ramnauth et al., 2022	T	Exp.|Non-RCT	Cog.	Home	Adult	Jibo	A	Trainer	Dyadic
[269]	Ranatunga et al., 2012	T	Exp.|Non-RCT	Im., Motor, Com.|NV	Clinic	Middle	Zeno	S	Trainer	Dyadic

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[209]	Robins et al., 2004	T	Exp.|Non-RCT	Gaze, Im., TT	School	Mixed	Robota	A	Peer	Triadic|Peer
[210]	Robins et al., 2004	I	Exp.|Non-RCT	Gaze, JA, Motor, Com.|NV	School	Mixed	Robota	A	Peer	Triadic|Ther.
[267]	Robins et al., 2005	I	Exp.|Non-RCT	Gaze, Im., JA, Motor, TT	School	Mixed	Robota	A	Peer	Triadic|Peer
[271]	Robins et al., 2009	T	Obs.|CS	Gaze, Im., JA, Motor	School	Mixed	KASPAR	S	Mediator	Triadic|Ther.
[275]	Robins et al., 2014	T	Case or SS	Cog., Im., JA, TT	School	Mixed	KASPAR	S	Peer	Triadic|Ther.
[777]	Robles-Bykbaev et al., 2018	I	Other|*unk.*	Cog.	School	Mixed	Custom	S	Trainer	Triadic|Ther.
[778]	Rodríguez-Quevedo et al., 2023	T	Case or SS	Emo., Eng.	School	Mixed	NAO	S	Trainer	Triadic|Peer
[779]	Romero-García et al., 2021	T	Other|*unk.*	Gaze, Motor, Com.|NV	Clinic	Early	NAO	S	Trainer	Triadic|Ther.
[780]	Rudovic et al., 2017	T	Other|*unk.*	Cog., Emo., Im., Eng.	Clinic	Mixed	NAO	A	Trainer	Triadic|Ther.
[781]	Rudovic et al., 2018	T	System	Cog., Emo., Eng.	Clinic	Mixed	NAO	S	Other|G	Triadic|Ther.
[782]	Saadatzi et al., 2018	C	Other|*unk.*	Cog.	School	Adult	NAO	S	Peer	Triadic|Peer
[783]	Saadatzi et al., 2018	C	Other|*unk.*	Cog.	Clinic	Early	NAO	S	Peer	Triadic|Other
[784]	Saha et al., 2021	T	Other|*unk.*	Cog., Im., JA, TT	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.
[785]	Salter et al., 2006	T	Obs.|CS	Eng., Motor	Lab	Early	Pekee	A	Other|G	Dyadic
[786]	Salvador et al., 2015	T	Other|*unk.*	Emo.	Lab	Mixed	Zeno	S	Trainer	Dyadic
[787]	Sandygulova et al., 2019	T	Other|*unk.*	Emo., Im., TT	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.
[788]	Sandygulova et al., 2022	T	Obs.|CS	Emo., Emo., Im., Eng.	Clinic	Mixed	NAO	S	Peer	Triadic|Ther.
[789]	Santos et al., 2020	I	Other|*unk.*	Cog., Im., Motor, Com.|NV	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.
[790]	Santos et al., 2022	T	Other|*unk.*	JA	Clinic	Early	NAO	S	Peer	Triadic|Ther.
[3]	Scassellati et al., 2018	T	Case or SS	Cog., Emo., JA	Home	Mixed	Jibo	A	Trainer	Triadic|Parent
[791]	Schadenberg et al., 2020	T	Other|*unk.*	Com.|V	School	Mixed	Zeno/Milo	S	Trainer	Triadic|Ther.
[792]	Schadenberg et al., 2021	T	Other|*unk.*	Cog., Gaze, Eng.	School	Mixed	Zeno	S	Trainer	Triadic|Peer
[257]	Schreider et al., 2024	I	Exp.|Non-RCT	Motor	Other|*unk.*	Mixed	MARIA T21	S	Trainer	Triadic|Ther.
[287]	Shahverdi et al., 2023	T	Exp.|Non-RCT	Gaze, Motor, Com.|NV	Clinic	Adult	Furhat	S	Trainer	Dyadic
[793]	Shamsuddin et al., 2012	I	Case or SS	Stereotypy, Gaze, Eng.	Clinic	Mixed	NAO	A	Trainer	Triadic|Peer
[794]	Shamsuddin et al., 2012	T	Obs.|CS	Gaze, Eng., Im., Motor, Com.|NV	Clinic	Mixed	NAO	S	Trainer	Triadic|Peer
[268]	Shamsuddin et al., 2012	I	Case or SS	Emo., Motor, Com.|NV	Clinic	Mixed	NAO	A	Trainer	Dyadic

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[795]	Shamsuddin et al., 2013	T	Other|*unk.*	Gaze, Eng., Motor, Com.|NV, Com.|V	Clinic	Mixed	NAO	S	Trainer	Triadic|Ther.
[796]	Shamsuddin et al., 2014	T	Exp.|Non-RCT	Eng.	Clinic	Mixed	NAO	A	Trainer	Triadic|Peer
[797]	She et al., 2018	T	Other|*unk.*	Com.|V	Clinic	Middle	LEO	S	Trainer	Dyadic
[798]	She et al., 2021	I	System	Com.|V	Lab	Mixed	NAO	S	Trainer	Dyadic
[799]	Shi et al., 2022	T	Exp.|Non-RCT	Eng.	Home	Early	Custom	A	Other|G	Dyadic
[800]	de Silva et al., 2009	T	Other|*unk.*	Gaze, Im., JA	Clinic	Mixed	HOAP-3	S	Trainer	Triadic|Peer
[801]	de Silva et al., 2009	T	Other|*unk.*	Gaze, JA, Motor, Com.|NV	Clinic	Mixed	HOAP-3	S	Trainer	Triadic|Peer
[297]	Silva et al., 2019	I	Exp.|Non-RCT	Emo., Gaze, Eng., TT	Clinic	Adult	Zoomer	A	Peer	Triadic|Ther.
[802]	Silva et al., 2024	T	Exp.|Non-RCT	Cog., Im., Motor, Com.|NV	School	Middle	NAO	S	Trainer	Triadic|Ther.
[803]	Silvera-Tawil et al., 2018	T	Case or SS	Com.|V	School	Teen	NAO	S	Peer	Other
[804]	Simut et al., 2016	C	Exp.|Non-RCT	Emo., Gaze, JA	School	Early	Probo	S	Trainer	Triadic|Ther.
[805]	Simut et al., 2016	C	Case or SS	Stereotypy, Com.|V	Clinic	Middle	Probogotchi	A	Peer	Dyadic
[806]	Singh et al., 2023	T	Case or SS	Other|*unk.*	School	Mixed	Custom	A	Trainer	Dyadic
[807]	So et al., 2018	C	Exp.|RCT	Motor, Com.|NV	School	Early	NAO	A	Trainer	Dyadic
[808]	So et al., 2018	C	Exp.|RCT	Com.|NV	School	Middle	NAO	A	Trainer	Dyadic
[338]	So et al., 2019	C	Exp.|RCT	Cog., JA	Other|*unk.*	Early	NAO	A	Trainer	Triadic|Other
[809]	So et al., 2019	C	Exp.|RCT	Cog., Motor, Com.|NV	Other|*unk.*	Early	NAO	A	Trainer	Triadic|Other
[810]	So et al., 2019	C	Exp.|RCT	Com.|NV	School	Middle	NAO	A	Trainer	Dyadic
[811]	So et al., 2023	C	Other|*unk.*	Cog., JA	School	Early	HUMANE	S	Trainer	Dyadic
[812]	So et al., 2023	C	Exp.|RCT	Cog., JA	School	Middle	HUMANE	S	Trainer	Dyadic
[813]	Soares et al., 2019	I	Case or SS	Emo., Emo., Im.	School	Middle	ZECA	S	Trainer	Triadic|Ther.
[305]	Soleiman et al., 2014	T	Other|*unk.*	Cog., Gaze, Com.|V	Clinic	Early	RoboParrot	S	Peer	Triadic|Ther.
[299]	Soleiman et al., 2023	T	Case or SS	Emo.	Other|*unk.*	Middle	RoboParrot, Red	S	Peer	Triadic|Peer
[814]	Sperati et al., 2020	C	Pilot or Feasibility	Cog., Emo., Im., JA, Eng.	Clinic	Early	+me	S	Peer	Triadic|Parent
[264]	Srinivasan et al., 2013	C	Exp.|Non-RCT	Im., Motor	Lab	Early	Isobot	A	Trainer	Dyadic
[475]	Srinivasan et al., 2015	C	Exp.|RCT	Stereotypy, Emo.	Clinic	Middle	NAO, Rovio	S	Trainer	Triadic|Peer
[815]	Srinivasan et al., 2016	C	Exp.|RCT	Cog.	Lab	Middle	NAO, Rovio	S	Other|G	Triadic|Ther.
[816]	Stanton et al., 2008	T	Exp.|Non-RCT	Eng., Com.|V	Lab	Middle	AIBO	A	Peer	Triadic|Other
[817]	Straten et al., 2018	T	Exp.|Non-RCT	Eng.	Other|*unk.*	Early	NAO	S	Peer	Dyadic

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[235]	Stribling et al., 2009	C	Case or SS	Eng.	Clinic	Middle	Labo-1	A	Other|G	Dyadic
[818]	Suzuki et al., 2022	T	Obs.|CS	Emo., Im.	Clinic	Teen	NAO	A	Peer	Dyadic
[819]	Syrdal et al., 2020	T	Longitudinal	Emo.	Other|*unk.*	Early	KASPAR	S	Peer	Triadic|Peer
[820]	Taheri et al., 2018	I	Case or SS	Gaze, Im., JA, Com.|NV, Com.|V	Clinic	Early	NAO, Alice-R50	S	Peer	Triadic|Ther.
[821]	Taheri et al., 2018	T	Case or SS	Cog., Im., JA, Com.|NV, Com.|V	Other|*unk.*	Early	NAO, Alice-R50	S	Trainer	Triadic|Peer
[262]	Taheri et al., 2020	T	Exp.|Non-RCT	Cog., Im., JA, Motor	Clinic	Early	NAO	N	Trainer	Triadic|Peer
[822]	Taheri et al., 2021	T	Exp.|Non-RCT	Cog., Im., JA, Eng.	Other|*unk.*	Early	NAO	S	Trainer	Triadic|Parent
[301]	Takata et al., 2023	C	Exp.|Non-RCT	Emo.	Lab	Teen	A-Lab ST, CommU, Sota	S	Peer	Triadic|Other
[823]	Talaei-Khoei et al., 2017	T	Case or SS	Cog., JA, TT	Lab	Middle	NAO	S	Peer	Dyadic
[824]	Telisheva et al., 2022	T	Other|*unk.*	Cog., Emo., Im., JA, Eng., Com.|V	Clinic	Early	NAO	S	Peer	Triadic|Parent
[824]	Telisheva et al., 2022	T	Other|*unk.*	Eng., Com.|V	Facility	Early	NAO	S	Trainer	Dyadic
[825]	Tleubayev et al., 2019	I	Case or SS	Gaze, Im., Eng.	Facility	Middle	NAO	S	Peer	Dyadic
[345]	Trombly et al., 2022	T	Exp.|Non-RCT	Emo., Eng.	Clinic	Early	Pepper	N	Trainer	Other
[826]	Valadão et al., 2016	I	Other|*unk.*	Gaze, Im., Eng., Motor, Com.|V	Other|*unk.*	Early	Custom	N	Peer	Dyadic
[827]	Vanderborght et al., 2012	I	Case or SS	Eng.	Other|*unk.*	Early	Probo	N	Trainer	Triadic|Ther.
[828]	Villano et al., 2011	T	Case or SS	Eng.	Lab	Middle	NAO	N	Peer	Triadic|Ther.
[829]	Wainer et al., 2010	T	Exp.|Non-RCT	Eng.	School	Middle	KASPAR	A	Peer	Dyadic
[278]	Wainer et al., 2014	T	Exp.|Non-RCT	Emo.	School	Middle	KASPAR	A	Trainer	Triadic|Peer
[830]	Wan et al., 2019	T	Other|*unk.*	Cog., Eng.	Facility	Early	Dabao, XiaoE, Mika	N	Peer	Triadic|Parent
[831]	Wanglavan et al., 2019	T	Pilot or Feasibility	Cog., Eng.	School	Early	BLISS	S	Trainer	Dyadic
[832]	Warren et al., 2015	C	Exp.|Non-RCT	Cog., Im.	Lab	Early	NAO	A	Trainer	Dyadic
[833]	Warren et al., 2015	C	Other|*unk.*	JA	Lab	Early	NAO	N	Trainer	Dyadic
[834]	Welch et al., 2023	T	Case or SS	Other|*unk.*	Other|*unk.*	Teen	NAO	N	Trainer	Dyadic
[207]	Werry et al., 2001	T	Case or SS	Cog., Eng., Im., JA	School	Middle	Custom	A	Peer	Triadic|Peer
[835]	Wong et al., 2016	T	Exp.|Non-RCT	Gaze, Im., JA, TT	School	Early	CuDDler	A	Trainer	Dyadic
[285]	Xie et al., 2024	T	Exp.|Non-RCT	Other|*unk.*	Lab	Teen	Pepper	A	Peer	Dyadic
[836]	Yaque et al., 2021	T	Exp.|Non-RCT	Eng., TT	Other|*unk.*	Early	Custom	S	Trainer	Dyadic
[837]	Yoshikawa et al., 2019	C	Other|*unk.*	Gaze	Lab	Teen	Actroid-F	N	Peer	Dyadic

#	Citation	Venue	Study Design	Targeted Skills	Location	Age Group	Robot	Operation	Role	Structure
[838]	Yun et al., 2014	T	Exp.|Non-RCT	Gaze, Im., Motor	Facility	Early	iRobi	A	Peer	Triadic|Ther.
[839]	Yun et al., 2016	T	Exp.|Non-RCT	Emo., Gaze	Facility	Early	iRobiQ, CARO	S	Peer	Triadic|Ther.
[840]	Yun et al., 2016	T	Exp.|Non-RCT	Emo., Gaze	Facility	Early	iRobiQ, CARO	S	Peer	Triadic|Ther.
[339]	Yun et al., 2016	C	Exp.|RCT	Emo., Gaze	Clinic	Early	iRobiQ, CARO	S	Trainer	Dyadic
[841]	Yussof et al., 2015	T	Exp.|Non-RCT	Stereotypy, Gaze, Im., Eng., Com.|NV, Com.|V	School	Middle	NAO	S	Peer	Dyadic
[842]	Zaraki et al., 2018	T	System	Cog.	School	Teen	KASPAR	S	Peer	Triadic|Other
[328]	Zaraki et al., 2020	T	Exp.|Non-RCT	Eng., TT	School	Mixed	KASPAR	A	Other|G	Triadic|Peer
[263]	Zhanatkyzy et al., 2023	T	Exp.|Non-RCT	Cog., Emo., Im., JA, Motor, Sensory	Other|*unk.*	Early	NAO	S	Peer	Dyadic
[843]	Zhang et al., 2019	T	Other|*unk.*	Other|*unk.*	School	Early	NAO	A	Other|G	Dyadic
[844]	Zhang et al., 2019	C	Other|*unk.*	Other|*unk.*	Lab	Early	NAO	A	Peer	Dyadic
[845]	Zheng et al., 2013	T	Exp.|Non-RCT	Gaze, JA	Clinic	Early	NAO	S	Trainer	Dyadic
[846]	Zheng et al., 2014	T	Exp.|Non-RCT	Cog., Im., Motor, Com.|NV	Clinic	Early	NAO	A	Trainer	Dyadic
[847]	Zheng et al., 2015	T	Exp.|Non-RCT	Cog., Im., Motor, Com.|NV	Lab	Early	NAO	A	Trainer	Dyadic
[847]	Zheng et al., 2016	T	Exp.|Non-RCT	Cog., Im., Motor, Com.|NV	Clinic	Early	NAO	A	Trainer	Dyadic
[848]	Zheng et al., 2018	T	Exp.|Non-RCT	Gaze, JA	Lab	Early	NAO	A	Trainer	Dyadic
[334]	Zheng et al., 2020	C	Exp.|RCT	Gaze, JA	Clinic	Early	NAO	A	Trainer	Dyadic