SOCIAL ROBOT INTERACTION DESIGN TO MITIGATE RISK IN SENSITIVE AND ADVERSE CONTEXTS

by

Terran Mott

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Robotics).

Golden, Colorado

Date ______________________

Signed: ______________________

Terran Mott

Signed: ______________________

Dr. Tom Williams
Thesis Advisor

Golden, Colorado

Date ______________________

Signed: ______________________

Dr. Andrew Petruska
Interdisciplinary Graduate Program Director
Department of Robotics

ABSTRACT

To be successful and acceptable, social robots must demonstrate social competence, navigate sensitive situations, and react to adverse events. Designing robot behaviors for these interactions is challenging because poor robot responses risk harming humans' dignity and well-being. This dissertation explores how social robots can be designed to effectively and appropriately respond to adverse or sensitive social interactions in positive ways that minimize risk to users' well-being. Chapter 2 begins by exploring an instance in which social robots are already used in the wild for potentially sensitive interactions— the use of teleoperated socially assistive robots in education, therapy, and telehealth for children. This work demonstrates the advantages of human oversight in this domain by identifying users' existing strategies to mitigate the social and emotional risks of child-robot interaction. It then presents design recommendations summarizing how roboticists can develop tools that support users' ability to prepare for and adapt to unforeseen situations.

Chapters 3 and 4 evaluate interaction design for autonomous robots in adverse interactions involving norm violations, such as unethical commands or hate speech. Chapter 3 explores how people appraise these interactions and investigates why they may prefer a robot to intervene or abdicate from responding to adverse events. Chapter 4 furthers this work through an empirical evaluation of robots' use of human-like linguistic politeness cues to address unethical commands. It presents a framework delineating how robots could use human-like cues to effectively and appropriately address adverse interactions while avoiding negative perceptions. This work also reemphasizes broader concerns about the extent to which robots should be able to perceive and react to such scenarios. Overall, this dissertation makes empirical and design contributions to the field of HRI that inform how social robots can preserve humans' dignity and well-being in adverse interactions. It argues that these contexts require roboticists to recognize factors outside of individual human-robot interactions— including the experiences of secondary stakeholders and bystanders, existing sociocultural norms of collaboration and conflict, and the potential for ill use of robots' capabilities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First I must share my profound gratitude for my incredible advisor, Dr. Tom Williams. Tom invited me into his lab with no formal computer science training and supported me through every spontaneous project I pivoted to throughout my PhD. I am a better writer, researcher, and leader for having the privilege of Tom's guidance and kindness in my life. From Tom, I have learned the discipline and emotional fortitude to turn mere good ideas into truly good work and to approach all things with integrity. I am sure that my future path involves more spontaneous pivots to new projects, but I will carry these skills with me wherever I go.

I also wish to thank all the other members of my committee: Dr. Katie Winkle, who took me in as a lab visitor at Uppsala and welcomed me to the amazing HRI community. Dr. Estelle Smith, for her guidance helping me see a light at the end of the tunnel. Dr. Elizabeth Reddy, who was a consistent source of steadiness and insight through the process of my PhD and dissertation. And Dr. Andrew Petruska, who believed in my ability to jump into grad engineering classes that became some of the coolest and most empowering experiences of my time at Mines.

This experience would not have been so special without the energy, creativity, and support of my labmates—Alexa, Rena, Puck, Rafa, Mark, and Saad—who are the best group of people anyone has ever done a PhD with. It has been fantastic to share meals, traditions, and travels with such a trustworthy and compassionate group of people.

I also must express my sincere thanks and gratitude to my parents, who have given me the support to invest in myself and my future through this PhD. It is because of them that I see joy in learning about the world and hold the self-belief to embrace new challenges with enthusiasm.

Finally, thank you to Michael, for his loyalty and support in the most challenging seasons of my PhD that went beyond anything words of encouragement alone could offer. And to all our friends for their continued support: thank you to Anna, Julia, Pete, Joe, Lynn, Henry, and Jeman. I can't wait to see you all in the city!

CHAPTER 1
INTRODUCTION

Robots of the past were almost exclusively built for industrial manufacturing. They performed repetitive tasks requiring few interactive or social capabilities. These constrained tasks often involved only *closed world* actions, in which there was no need to perceive nor adapt to unforeseen changes in the environment. Robots of the future, however, will have far broader and more complex responsibilities. The implementation of robots outside of industrial settings will require that they succeed in far more unpredictable and unconstrained environments than an assembly line. Current and near-future robots must be far more versatile and interactive than their industrial ancestors—able to physically and socially engage with people in everyday places.

Robots' potential to interact with humans outside of manufacturing creates new opportunities for new user communities. Social robots can decrease human burdens and add value to human experiences in many familiar domains. For example, robots can provide assistance in care facilities and hospitals by relieving humans of tiring, routine tasks [1, 2], as well as provide comfort, entertainment, or companionship [3–5]. Robots in education can provide personalized, non-judgemental instruction [6–9] and equip educators with new interactive communication tools [10–12]. Similarly, robots can offer assistance or entertainment in the home [13, 14] and in a variety of public spaces [15]. Even in industrial manufacturing, robots' new capabilities open up possibilities for human collaboration [16].

Robots developed for these everyday settings become part of their social environment as well as their physical environment. Some robots will be explicitly designed for socialization, but many others will engage socially in service of practical tasks. Regardless, robots' acceptability and success in the wild will often depend on their ability to integrate with complex social practices [17, 18]. Humans expect social robots to carry the competencies and obligations of a social peer [19], meaning that robots must participate in the social norms of humans around them to gain acceptance. For instance, robots use social skills when they respect humans' personal space [20, 21], when they use polite language [22, 23], or when they give humans feedback and advice [24]. Robots also take on familiar social roles, such as tutors [8, 9], teammates [25], or companions [3]. Ultimately, social robots' ability to competently engage with the social conventions of their user communities influences whether they are worthy of acceptance [26, 27] and trust [28–30].

The interdisciplinary field of Human-Robot Interaction (HRI) seeks to understand the effects of social robots' presence, to analyze the benefits and burdens that robots may bring to user communities, and to design robots for socially and ethically beneficial outcomes. HRI draws on interdisciplinary work in robotics, cognitive science, linguistics, and philosophy to explore the complex sociotechnical ramifications of social robots' design and deployment [18, 29, 31, 32]. HRI research prioritizes understanding human perspectives on social robots' actions and abilities. It often involves open-ended engagement with stakeholders to understand what ought to be designed (or not designed) to meet the needs and constraints of current or potential robot user communities. This type of work is often accompanied by human-subjects studies that evaluate potential robot behaviors to determine which may be successful.

1.1 Social Robots Introduce Distinct Risks

A key responsibility of the HRI research community is to comprehend and address the distinctive risks that social robots introduce to user communities. As participants in their social environment, robots will encounter unexpected, adverse events that are bound to occur in social interaction. Social robots are already being developed for domains in which these sensitive interactions are probable: where they assist vulnerable user communities or aid users in serious circumstances. For instance, social robots can engage with children [33, 34], support older adults [22, 35], and offer novel communication tools for people with complex communication needs [36–38]. Similarly, robots can support people in a variety of care domains by helping both patients [3] and caregivers [1, 2, 39]. Robots can provide assistance or encouragement in rehabilitation [40, 41], in various therapeutic settings [11], and in mental health services [42, 43]. Robots in these sensitive domains will inevitably partake in interactions that bear risk to users' well-being [44–46]. Their behaviors have the potential to strengthen or undermine social relationships [47], to perpetuate or mitigate stigma [45], and overall to support or harm humans' dignity [48].

Social robots will also encounter adverse social interactions when people do something wrong or inappropriate. Robots that collaborate with humans will witness and possibly join in human conflict and conflict resolution [25, 49, 50]. Robots will have the opportunity to blame humans or to bear the blame for mistakes and failures [51–53]. They may advise humans in ethical decision-making or provide input during tense decisions [53, 54]. Robots will also receive rude or abusive behavior from humans [55] or witness humans' rude or abusive behavior as bystanders [48, 56]—including instances of bias or prejudice. Humans will give robots unethical commands that directly request harmful actions [57] or request a robot's complacency in them [58]. There are considerable risks that inappropriate or insufficient responses from robots in these situations will harm the humans involved. Poor robot response behaviors may perpetuate damaging stereotypes [59, 60] or reinforce bias [56]. Even a robot's "non-reaction" of ignoring a situation may implicitly condone unethical actions or weaken humans' perception of moral norms [61, 62]. Robots must be designed to engage appropriately and effectively in these situations to mitigate the potential for social or emotional harm to users and bystanders. They must identify morally fraught interactions [63, 64] and use effective communication strategies to address them [24, 57] in ways that preserve the dignity and well-being of humans involved.

1.2 Open Questions Remain About How Robots Should Be Designed to Engage in Sensitive Interactions

Many open questions remain in the HRI community regarding how robots ought to be designed to effectively address adverse, socially sensitive interactions. To explore these questions, roboticists can work to understand user communities' values and concerns to comprehend *why* they may expect or prefer various robot behaviors. In this way, roboticists can identify the features of adverse interactions that should inform a robot's behavior or language. Critically, roboticists must also evaluate whether robots should involve themselves at all, and consider when it may actually be harmful for robots to initiate or intervene in an adverse interaction.

To investigate these questions, roboticists and interaction designers must recognize that not all social robots are autonomous—robots have a wide variety of Levels of Autonomy across domains. Level of Autonomy (LoA) describes the extent to which a robot's actions are monitored, controlled, or approved by humans [65–67]. In this way, robot autonomy is represented by a complicated spectrum of design choices, rather than a binary distinction between low- and high-autonomy systems. For instance, a robot capable of many autonomous behaviors may necessarily seek human approval for task-critical actions. A robot that requires significant human oversight might still autonomously generate many potential actions and require a human simply to select the best one. These autonomy-based design choices have a significant impact on a robot's success [68, 69] and affect how credit or blame is allocated when it fails [70]. Though visions of high-LoA robots pervade media and motivate much HRI research, many robotic systems in the wild challenge this assumption. While some have significant autonomy [1, 8], many others rely heavily on human oversight in social interaction [11]. A robot's LoA fundamentally changes the challenges facing roboticists studying effective robot behaviors in adverse, sensitive interactions. Therefore, understanding the role of human oversight in different domains is a key first step to designing robots that can address adverse events in positive ways.

1.3 How Can Roboticists Support Users of Low-Autonomy Robots in Sensitive Interactions?

For robots on the lower end of the spectrum of Levels of Autonomy [66], a human user's ability to observe and adapt to adverse events is key to a robot's ability to respond appropriately and effectively. However, there are many potential trade-offs involved with reliance on human decision-making. For instance, people who monitor or operate robots must maintain Situation Awareness of both the robot's control interface and the interaction at hand. These cognitive demands can impact operators' ability to notice problems and adapt effectively [71, 72]. For this reason, some researchers have argued that low-autonomy robots may be undesirable in socially sensitive domains [73]. However, others have advocated for low-autonomy social robots. These researchers propose that robots in highly unpredictable domains may require human oversight to minimize the risk of negative impacts on humans involved [74]. The potential impacts of these trade-offs on robots' ability to perform beneficial actions in sensitive interactions represent critical, yet underexplored questions in HRI research.

It is also often unclear how roboticists can best support the unique set of users and secondary stakeholders involved in low-autonomy HRI domains. Many low-autonomy social robots are currently used in socially assistive settings, such as therapy, healthcare, and education. In these domains, the robot operator (usually a care provider or other expert) differs from the person benefiting from the robot interaction (a client or patient). *Operators* represent a distinct category of end users, in addition to the people a robot was designed to assist. Previous HRI research focusing on lab studies of dyadic interactions with assistive robots has neglected to consider the needs and experiences of these users. This perspective ignores the role of robot operators in adopting robots, adapting to their limitations, and assessing their trustworthiness. Therefore, HRI researchers must explore the development of robots and their control interfaces that meet the needs of operators, as well as assisted individuals. In particular, roboticists can work to understand the strategies robot operators already use to address unexpected or adverse interactions to minimize the risk of negative impacts on others.

1.4 When and How Should Autonomous Robots React to Sensitive Interactions?

Not all robots can rely on human operators' oversight. Many current and near-future robots will confront adverse situations that they must react or respond to autonomously. In these cases, roboticists face a different set of design challenges. In particular, roboticists face open questions about whether or when robots should initiate or intervene in fraught interactions at all. Robots may not always possess the social standing to engage in sensitive situations because while robots are social participants, understanding their status as social peers is rarely straightforward. They occupy a unique ontological niche [75]—they are inanimate, yet many can perform and react to social actions [76] and take on human-like social roles [24, 63]. Robots may merely depict or perform human social abilities [77] without truly possessing them. For these reasons, people perceive robots as "social others" without the same social, moral, or emotional standing as other humans [78]. They often expect robots to have less social power and more relational distance compared to humans in equivalent roles [79, 80]. Robots' fundamentally limited status means that they may not always possess the standing to partake in vulnerable conversations or to take potentially threatening social actions [81]. In many situations, it may not be suitable for a robot to involve themselves in rebukes, criticism, or conflict with humans. Therefore, an open question facing the HRI community is under which conditions robots possess the social standing to react to an adverse event at all—roboticists must explore when people feel that social robots ought to intervene or disengage in fraught interactions and why.

Furthermore, when social robots do autonomously intervene in an adverse or sensitive interaction, it is unclear whether they ought to do so in the same way humans might react to an equivalent situation. Previous HRI research has established complex trade-offs related to this interaction design question. On the one hand, social robots that mimic human social behaviors are often successful: robots that follow human norms are perceived as more socially competent [26, 27], predictable, and acceptable [22, 82–84]. On the other hand, human-like behavior can backfire when used by robots. Human norms do not always apply to robot interactions [19, 85]. It can be uncanny, deceitful, or disingenuous for robots to mimic human linguistic and social cues [86–89]. It is unclear under which circumstances human-like language might help or hurt robots' acceptability and social competence. By evaluating these trade-offs, roboticists can develop an understanding of whether robots ought to mimic humans when reacting to an adverse or unexpected part of social interaction.

1.5 Contributions of this Dissertation

The main chapters of this dissertation (Chapters 2, 3, and 4) present three interrelated research projects that explore open questions about how social robots ought to be designed to address sensitive, adverse interactions in positive ways. Each project considers a distinct perspective on this multifaceted challenge and investigates specific research(s) questions within it. While these research questions are only some of the many considerations that arise in designing sensitive robot behaviors, they represent a consistent thread of investigation that contributes a foundation for further HRI research on this topic. The first project presented in Chapter 2 explores research questions regarding how to support users of low-autonomy social robots, which require significant human oversight and control. This work studies the use of socially assistive robots in assistive domains with children—a real-world deployment context in which robots are already used for sensitive interactions in the wild. In this way, the project presented in Chapter 2 reflects an assumption that robots in a given context necessarily have low autonomy. Based on this constraint of studying low-autonomy robots and their user communities, Chapter 2 investigates the design of robot systems and user interfaces that empower users to enact positive robot interactions when unexpected or adverse events occur.

The research presented in Chapter 2 is constrained to the study of low-autonomy robotic systems in the wild. Alternative constraints lead to parallel research questions based on the assumption that robots in sensitive interactions could instead have relatively high autonomy. Therefore, Chapters 3 and 4 consider this parallel premise—investigating the case in which robots react to adverse social interactions with little immediate human oversight. Because this work is not tied to a specific real-world user community, the projects presented in these chapters utilize different research methods, including both online surveys and lab-based experiments. First, the work presented in Chapter 3 investigates research questions regarding whether robots should engage in vulnerable or threatening interactions, how people assess such robot behaviors, and why some may be perceived more positively. This project takes an exploratory perspective to establish situational and contextual factors that impact the success of autonomous robots in adverse interactions. Finally, the third project presented in Chapter 4 further considers the assumption that robots with a high degree of autonomy may need to generate response behaviors in fraught scenarios. Compared to Chapter 3's exploratory perspective, Chapter 4 considers much more pointed research questions aimed at evaluating robots' use of specific linguistic behaviors based on human sociolinguistics literature. Specifically, Chapter 4 investigates research questions regarding the type of linguistic cues that are most appropriate and effective for robots in adverse interactions.

In this way, the questions considered in Chapters 2, 3, and 4 represent a consistent line of research across domains and Levels of Autonomy. Chapter 2 investigates a real-world case where robots have low autonomy in sensitive interactions. Chapter 3 explores an alternative case, in which robots navigate adverse interactions with no human oversight. Finally, Chapter 4 furthers this exploratory work through a specific evaluation of robots' use of various linguistic strategies. Overall, the work discussed through these three projects makes empirical and design contributions through qualitative and quantitative projects across multiple domains. I find that developing ethical, effective robot behaviors in adverse interactions requires careful consideration of robot autonomy and linguistic anthropomorphism, as well as of whether robots should engage in sensitive interactions at all. Through these contributions, I demonstrate that mitigating social or emotional risks in potentially adverse human-robot interactions requires roboticists to consider broader situational factors —such as secondary stakeholder expertise, the role of bystanders, or the potential for robots to be misused beyond the scope of their intended purpose.

1.5.1 Empirical Contributions

This dissertation presents findings from empirical evaluations of robotic systems' responses to adverse interactions across domains and Levels of Autonomy. Chapter 2 begins this dissertation by exploring an instance in which social robots are already used consistently in the wild in potentially risky interactions—the use of teleoperated socially assistive robots in education, therapy, and telehealth for children. In this low-autonomy use case, I provide evidence for the advantages of human oversight by identifying therapists' existing strategies to mitigate the social and emotional risks of child-robot interaction in therapeutic settings. Chapter 2 provides a significantly improved understanding of how robotic tools can meet the institutional and individual needs of user communities in this domain. This broader perspective demonstrates how consideration for the needs of secondary stakeholders is key to providing the most benefit to children themselves while minimizing the risk of social or emotional harm.

Because not all robots can be teleoperated in adverse interactions, Chapters 3 and 4 evaluate interaction design for autonomous robots through a focus on norm-sensitive robotic noncompliance. These projects investigate how robots should be designed to respond when they receive a norm-violating command or when they witness norm violations, such as hate speech. Noncompliance interactions are particularly sensitive because robots' actions have significant potential to affect humans in positive or harmful ways. Chapter 3, a qualitative survey, presents evidence showing *why* people may perceive particular robot response strategies as effective or inappropriate. Furthermore, Chapter 3 contributes an understanding of the key situational and psychological factors that determine whether robots should intervene or react to adverse interactions at all. It argues that existing sociocultural norms of collaboration determine whether people expect robots to bear responsibility for responding to norm violations or whether they expect robots to abdicate such behaviors to other humans.

When robots do intervene to respond to an adverse event, there is a wide range of linguistic cues that they can potentially use to generate tactful responses. Therefore, Chapter 4 presents an empirical evaluation of robots' use of human-like linguistic strategies in noncompliance responses to unethical commands. Chapter 4 provides evidence that while people expect robots to act tactfully, they do not expect them to strictly mimic human linguistic behaviors. It then presents a framework delineating how robots can use human-like cues to respond competently to adverse interactions while avoiding negative perceptions.

1.5.2 Design Contributions

This dissertation provides theoretical and practical insights for interaction design to minimize social or emotional harm to robot users in adverse, sensitive interactions. In Chapter 2, a low-autonomy domain, I present design recommendations summarizing how roboticists can design assistive robot therapy tools that support users' ability to adapt to unforeseen situations. For higher-autonomy domains, Chapter 3 presents design considerations that can inform whether robots should be designed to initiate or intervene in adverse interactions when collaborating with humans. Finally, Chapter 4 contributes explicit design recommendations for the types of linguistic cues that robots can use to offer effective, appropriate norm-violation response behaviors through the framework of *bounded proportionality*, in which robots are limited to linguistic cues that avoid inappropriate human-likeness.

1.5.3 User-Centered Contributions

Qualitative findings from this dissertation contribute a nuanced picture of the values, needs, and potential concerns of current and near-future robot users. In Chapter 2, I present an analysis of the needs and practices of teleoperated socially assistive robot users through in-depth interviews with early adopters. Qualitative results from Chapters 3 and 4 emphasize the values and assumptions people use to appraise social robot behaviors and show why particular norm-sensitive actions may be successful or unsuccessful. In particular, Chapter 4 emphasizes peoples' concerns about whether robots should be capable of surveilling or rebuking human behavior. In this way, Chapter 4 raises larger questions about the extent to which robots should be able to perceive and react to such scenarios.

1.6 Conclusion

This dissertation demonstrates that successful human-robot interactions in adverse, sensitive situations require interaction designers to confront key considerations about robots' autonomy, human likeness, and social roles. Chapter 5 concludes this dissertation by summarizing key results from the proceeding chapters. It emphasizes that social robots' ability to mitigate social or emotional risks to user communities depends on roboticists' understanding of contextual factors outside of individual robot interactions. Finally, Chapter 5 explores avenues for future work on this topic focused on how roboticists can prioritize transparency in this interaction design space and support robot user communities more broadly.

CHAPTER 2
SUPPORTING EARLY ADOPTERS & NOVICE TELEOPERATORS OF SOCIALLY ASSISTIVE ROBOTS

Modified from the following papers:

A paper published at the ACM Conference on Computer Supported Collaborative Work and Social Computing in 2023. Saad Elbeleidy[1], Terran Mott[2], Dan Liu[3], Ellen Do[4], Elizabeth Reddy[5], and Tom Williams[6].

A paper published at the IEEE Conference on Robot and Human Interactive Communication in 2022. Saad Elbeleidy, Terran Mott, Dan Liu, and Tom Williams. © 2022 IEEE

A paper published in alt.HRI at the ACM/IEE Conference on Human-Robot Interaction in 2022. Saad Elbeleidy, Terran Mott, and Tom Williams. © 2022 IEEE

The textual material included in this chapter is originally written by Terran Mott.

2.1 Motivation

As robots' social skills develop, so too does their ability to provide benefits to users in social interactions. Socially assistive robots (SARs) are robots that are explicitly designed to provide assistance through social interaction [90] rather than through physical actions. SARs have shown significant potential to serve potentially vulnerable user communities—such as in care domains. SARs can support older adults through companionship [35] and assist them in living independently [22, 91]. SARs also have significant potential as tools to provide therapeutic interventions, assistance, or encouragement in rehabilitation and physical therapy [40, 41, 92] and in mental health services [42, 46]. Although social assistive robots have shown significant potential in these sensitive domains, it is uncertain how to support their success for long periods in the wild. Roboticists must assess the benefits and risks of assistive robots, as well as work to

[1] Graduate student, Colorado School of Mines

[2] Graduate student, Colorado School of Mines

[3] Graduate student, University of Colorado Boulder

[4] Director of Partnerships and Innovation, ATLAS Institute, University of Colorado Boulder

[5] Assistant Professor, Colorado School of Mines

[6] Assistant Professor, Colorado School of Mines

understand how to design systems that support both assisted individuals and secondary stakeholders within their user communities.

2.2 Socially Assistive Robots Introduce Both Benefits and Risks for Children

Socially assistive robots hold substantial potential to generate positive impacts through interacting with children in education, healthcare, and therapy [73, 93–95]. Social robots can successfully capture children's attention in unique ways compared to other potential interventions or other forms of interactive technology [12, 73]. Robots engage children in educational or therapeutic activities because they create interactive experiences without the threat of judgment or ostracism that may accompany interactions with adult authority figures or peers [9, 96]. In education, robots have the potential to enhance learning [12, 97, 98] by offering cooperative, playful experiences [8]. Social robots can also assist children in learning cognitive and collaborative skills [12, 73], such as self-reflection [99] and conflict resolution [49]. In therapeutic activities, robots can help provide a positive affective experience to children [41]. They can serve as rehabilitation tools for children with developmental disabilities [100], as well as assist in therapeutic interventions that decrease stress and increase children's well-being [101].

To responsibly investigate the potential of SARs in child-robot interaction, roboticists must also remain cognizant of specific risks associated with children as a vulnerable user population. Children are still developing an understanding of the world and naturally have less control over the technology in their lives [102, 103]. However, technologists working on assistive tools for children should avoid "deficit models" that frame children as cognitively incomplete [104] or that focus on reductive paradigms of "fixing" children [102]. Instead, researchers should approach children as intelligent stakeholders [104, 105] who require trustworthy, transparent technology [106]. Therefore, considering how social robots may adversely affect their right to autonomy, privacy, or truthfulness is critical [106, 107]. Adverse interactions between children and robots introduce particular risks to children's safety and well-being. For example, because children relate strongly to robots [108], there are distinct risks that children will develop problematic bonds with social robots [44, 109] and become prone to misplacing trust in them [110]. Children may not realize that robots are capable of deceptive self-description—such as a robot claiming to possess human-like emotions [44]. Similarly, children may develop inaccurate mental models about the extent to which robots are social or emotional beings and misunderstand that they are incapable of feeling pain [111] or keeping secrets [112]. Roboticists must work to mitigate these potential risks while exploring the many potential benefits of developing socially assistive robots for children.

2.3 Research on Assistive Robots for Children Focuses on Experimental Outcomes and Autonomous Robots

Because of robots' significant potential to promote beneficial outcomes for children, a wide variety of human-robot interaction research has explored SAR design for child-robot interaction. Much of this body of work relies on laboratory-based studies that identify the efficacy of robots by measuring educational or therapeutic outcomes after robot-based interventions. Studies have shown that robots can successfully assist in therapy by modeling or teaching communication skills [94], offering encouragement [93], or providing entertainment and generally increasing engagement [95]. In addition to focusing on children's measurable outcomes, much previous research on SARs in child-robot interaction advocates for autonomous robots. For example, some researchers have argued that fully autonomous robots can minimize the training burden and demands of operating robots for users with less technology expertise [90]. Others have argued that highly autonomous robots would be more feasible given the complexity of domains in which most SARs would be deployed [73]. These arguments cast doubt on the role of low-autonomy SARs [113], especially those that are fully teleoperated—completely controlled by an educator, therapist, or other practitioner. They suggest that the role of teleoperation in SAR research is merely to provide a tool for prototyping robot behaviors or running preliminary Wizard of Oz experiments [114]. As such, this work views teleoperation as a short-term solution that would be impractical or undesirable to incorporate in robots for deployment in the wild.

2.4 Socially Assistive Robots for Children Are Teleoperated in Practice

Experimental research has demonstrated the value of SARs in educational and therapeutic settings with children [73, 93–95]; however, lab studies do not capture how such robots are used in the wild. Many organizations have worked to make social robots (as shown in Figure 2.1) more widely available to therapists, educators, and other practitioners who work with children [10, 115–118]. These systems use inexpensive materials [10] or materials already available to users [115]. And in contrast to the assumptions of most SAR research, these robots are often teleoperated—therapists or educators control what they say and when. [11, 119, 120].

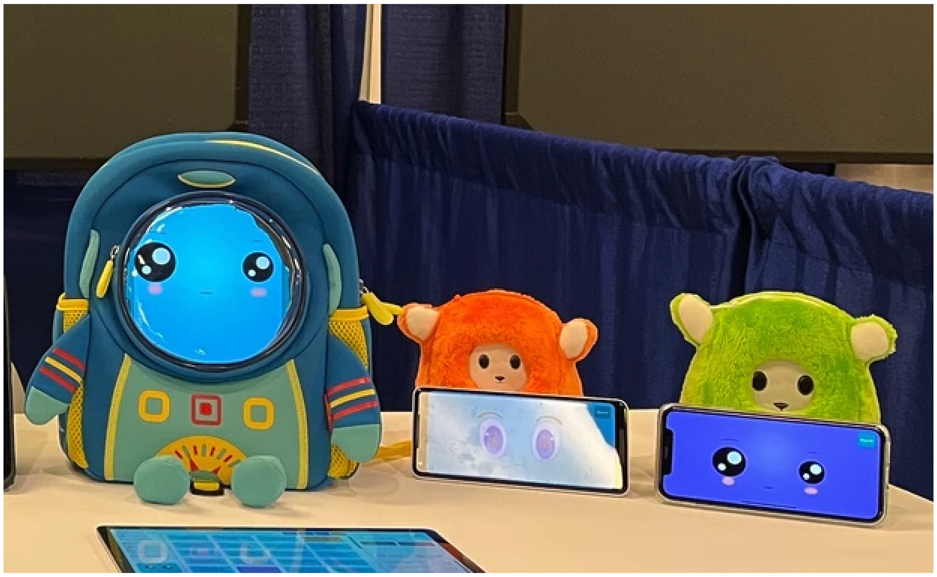


Figure 2.1 Socially assistive robots created using affordable materials.

There are several reasons why teleoperation has emerged as the dominant form of SAR control in real-world contexts. Teleoperation can reduce the cost and complexity of robot interactions, making robots more affordable and accessible to schools and clinics. These inexpensive robots are often safe and durable enough for children themselves [120]. Additionally, some HRI researchers have argued that teleoperated SARs may possess further benefits beyond their affordability and practicality [74]. They argue that low-autonomy robots can support practitioners' ability to identify and address potential problems [121] and maintain more awareness of the interaction at hand [74].

2.5 Teleoperators' Needs as the Primary Users of SARs Are Under-Explored

Robot teleoperators' needs have been mostly ignored in HRI research focused on individual child-robot interactions with autonomous robots [74]. Though therapists and educators have influenced the development of robotic technologies [40, 122–124], researchers have often involved them only as expert consultants in the development of autonomous systems, not as end users of the system itself. This perspective lacks a fundamental focus on adult practitioners as the *primary users* of SARs with children. It is these adult practitioners who are responsible for deciding to adopt SARs, for incorporating them into their existing practices, and for adapting to unexpected events in effective, emotionally positive ways. Yet it is not clear how SARs should be designed to fit into the complex environments of real-world schools and clinics, nor what barriers teleoperators may face when adopting SARs into their existing workflows. Additionally, it is unclear what strategies robot operators may already use to address sensitive or adverse events that inevitably occur outside of lab environments. To understand how SAR's can be successfully developed and deployed in the wild, researchers must center the needs, values, and concerns of teleoperators as primary users. Therefore, in this investigation, we ask the following research questions:

1. What are the values, needs, and existing practices of early adopters of SARs with children?
2. What challenges may new users face in adopting SARs?
3. How can SARs and their user interfaces be improved to support both familiar and novice users in child-robot interaction domains?

In this paper, we present the findings of two projects emphasizing the perspective of SAR teleoperators as the primary users of assistive robots in child-robot interaction, whose preliminary results were presented in [125] and [126]. First, we present the results of an interview study with *early adopters* who have already incorporated socially assistive robots into their practice with children. Based on these interviews, we characterize users' cyclical workflow and present six themes synthesizing the needs and practices of SAR teleoperators. Our interview findings demonstrate that centering teleoperators' perspectives reveals how robots must be designed to support their personal and institutional needs. On the one hand, robotic tools must support teleoperators' emotional awareness and ability to adapt quickly during robot interactions. At the same time, they must also support institutional and collaborative preparation and documentation practices that make their use possible. Second, we present the results of group usability tests with *novice users* who are expert therapists with no previous robot teleoperation experience. The results of these usability tests uncover further challenges to the adoption of SARs in the wild. We demonstrate that while novice users intuit the effectiveness of robots in therapy, they face a steep learning curve and bear concerns about the accessibility of therapeutic content and the effects of robotic therapy tools on client privacy.

These parallel projects highlight the values, needs, and risk-mitigation strategies of users who are already familiar with robots in this domain, as well as the potential challenges faced by new users adopting SARs for the first time. We conclude that it is not necessary to convince users to see the value of socially assistive robots for children. Instead, it is paramount to design robots that function in ways that are consistent with the broader institutional environment users work within. We argue that robotic technology that is sensitive to teleoperators' needs outside of individual child-robot interactions can be both more beginner-friendly and support long-term use.

We synthesize our findings into a set of design recommendations focused on supporting teleoperators in accessing the benefits of socially assistive robots while mitigating the risks of using robots with children as a vulnerable population. Our design recommendations focus on improving robot dialogue teleoperation technology in a way that centers operators (caregivers, educators, and therapists) as the primary stakeholders of this technology. Critically, our recommendations emphasize the value of low autonomy in this domain and argue that teleoperation should be considered a permanent design solution with both practical and ethical benefits.

2.6 Related Work

Socially assistive robots (SARs), which provide social rather than physical assistance, have the potential to positively impact many user communities. SARs can provide companionship to the elderly [35], support children in their education [12, 97, 127], or engage in therapeutic interventions [41, 42, 73, 101]. SARs are particularly effective in engaging with children [73, 128, 129]. From a child's perspective, robot interaction can be less intimidating than human interaction. Children often show high acceptance and likability for robots [5]. For these reasons, robots can be a particularly effective tool for adding interactivity and novelty to educational or social skills-focused activities with children. Dialog-based social robots can motivate children to interact, learn [130], and express their emotions [131, 132].

2.6.1 SARs in Therapeutic Domains with Children

The goals of therapy, in general, are to engage children to help them meet their needs for easier daily life [132]. Therapists choose from various methods to engage their clients, including music, dance, art, or social robots [132–134]. Various modalities of therapy focus on different life skills, including occupational therapy (OT) to improve movement, speech and language pathology (SLP) to improve speech and communication, and applied behavior analysis (ABA) therapy to improve social skills. Often, therapeutic interventions are intended to assist neurodiverse children—those with features including autism, attention-deficit / hyperactivity disorder (ADHD), dyslexia, Down's syndrome, bipolar disorder, and others [135, 136]. The neurodiversity movement, led largely by autistic self-advocates, has emphasized a social model for disability [137][7]. This is the idea that individuals neurological differences ought to be respected and appreciated, similarly to other individual differences in human variation [139]. The neurodiversity movement asks society to accept and support disabled individuals without an attempt to cure them [140]. This model of disability presents a perspective that de-pathologizes individuals while maintaining that neurological classification may still be useful and meaningful [141, 142].

The role of robots within the various therapeutic practices that are used to assist children is comparatively novel [143]. Researchers have demonstrated that socially assistive robots can successfully help neurodivergent children reach various learning goals, including collaboration [144], verbal expression [145–147], eye contact [147, 148], and play [149]. Many of these positive impacts on children's social skills have been shown also in long-term, longitudinal studies [150, 151]. A significant portion of Socially Assistive Robot research focuses on assistance for autistic individuals, particularly autistic children [73, 93–95].

Therapists themselves have conventionally influenced SAR research consultants in the development of autonomous systems [40, 122–124]. However, most currently available SAR platforms for children rely on teleoperation [10, 115]. Therefore, most therapists, educators, or other caregivers who use SARs in practice use these teleoperated robotic systems [11, 119, 120]. Social robot teleoperation requires tools specifically designed for dialogue teleoperation, in contrast to teleoperation tools for movement or manipulation. However, relatively little research exists on spoken dialogue authoring and teleoperation interfaces [11, 152, 153]. Though some similar tools do exist for virtual character authoring and game design, these fundamentally aim for autonomous content delivery, involve granular levels of behavior definition, and may not be

[7] The shift from a medical perspective to a social perspective is a core part of the change in language referring to neurodivergent individuals using identity-first language rather than person-first language [138].

accessible to educators and therapists whose expertise varies significantly from that of game designers and developers [154, 155].

2.6.2 Levels of Autonomy for Socially Assistive Robots

A key consideration for all robots in human environments is their Level of Autonomy. Beer et al. define autonomy as: “The extent to which a robot can sense its environment, plan based on that environment, and act upon that environment with the intent of reaching some task-specific goal without external control” [65]. This theory builds on the theory of levels of automation [66] by additionally considering robot capabilities, such as social interaction [65]. LoA selection guidelines suggest researchers examine three dimensions of the domain in which the robot will be deployed: task criticality, task accountability, and environment complexity. *Task criticality* is central to choosing a robot’s LoA because automation has direct consequences for task failure rates [156], and increased automation can introduce unique risks in the context of highly critical tasks [157, 158]. Even when autonomy is otherwise low, temporarily high autonomy may compromise an operator’s Situation Awareness [71]. Therefore, robots with less autonomy are recommended in domains with highly critical tasks. Similarly, domains with complicated *task accountability* may merit low autonomy. When robots are perceived as more autonomous, people give them more credit or blame for the outcomes of their actions [70]. For instance, users in safety-critical domains, such as clinical environments, have been reluctant to adopt automated technologies over liability concerns [159]. Finally, a task’s *environmental complexity* also influences the ideal LoA for robotic systems. Autonomous robots in complex, dynamic environments necessitate significant perceptual capabilities [160]. Even when a robot’s sensory capabilities are reliable, high-LoAs may only be effective when a complex environment is also predictable. In highly unpredictable environments, a robot may need to be teleoperated or, at minimum, supervised [121].

Due to the complex considerations that go into robot LoA design, there is disagreement about the appropriate level of autonomy for socially assistive robots. Much of the foundational work on SARs has been motivated by visions of a future with fully autonomous robots, or otherwise high LoAs. This perspective acknowledges the challenge of practically deploying teleoperated SARs and the perceived limitations of teleoperation interfaces [73, 90, 113]. In contrast, other perspectives in the HRI community have argued for implementing lower levels of autonomy in socially assistive domains to support caregivers through human augmentation rather than automation [161]. Our work contributes to increasingly important conversations about robots in therapy [73, 93, 94] with a new focus on the experiences of robot teleoperators.

2.7 Methods

The goal of this project was to investigate the needs and practices of early adopters of SAR teleoperators with children in the wild, as well as the challenges faced by novice users. To explore this topic, we conducted two parallel projects with different sets of participants. With experienced users, we conducted a set of semi-structured interviews. With novice users, we conducted a set of group usability tests exploring the experience of setting up and using a SAR for the first time.

2.8 The Peerbots Platform

We leveraged a collaboration with the assistive robotics nonprofit Peerbots [115] to facilitate both components of this project. Peerbots is a social robot software nonprofit that develops an interface that allows the teleoperation of robots in socially assistive robotics domains, from therapy to education. In our local context, this interface is used in Special Education classrooms, in which therapists use the Peerbots interface to control robots used as high-tech puppets in their classrooms. The Peerbots organization was instrumental in the recruitment of participants for both interviews and usability tests. Additionally, we used the Peerbots teleoperation interface (Figure 2.2) as a visual reference during interviews, and as a tool to simulate telehealth interactions during the group usability tests.

Yet while our collaboration with the Peerbots team (and their user community) helped to guide our research, our methods and findings are not specific to the Peerbots platform. A minimal portion of interviews and usability tests related to Peerbots-specific features, but the bulk of those interviews instead focused on participants' higher-level needs and workflows, which would likely be relevant to any assistive robotics teleoperation tool.

Here, we include a brief description of the Peerbots system in order to provide context for the methods employed in this project and to describe Peerbots' approach to developing a functional SAR platform in the wild. The Peerbots system is a versatile socially assistive robot teleoperation platform that relies on inexpensive materials. Peerbots requires two devices: either a phone and tablet or two tablets. A teleoperator holds one tablet, which serves as the robot dialog controller. This tablet displays “pallets”, sets of buttons that represent different things the robot can say or express. For instance, pallets might include the set of robot utterances required for the robot to greet a child or read a book to them. The other device is inserted into a soft, durable body (such as those in Figure 2.1). This device displays the robot’s face and provides its voice. When the teleoperator presses a button on their interface, it causes the robot-device to execute that utterance or expression. The controller- and robot-devices do not necessarily need to be co-located to connect to one another in this way. For this reason, the Peerbots platform can be used for co-located assistive interactions and for telehealth interactions. In telehealth interactions, a child and provider meet through an online video conference, with a robot body co-located with the child but still controlled by the provider. This allows the provider to have a more engaging, physical presence with the child, despite the challenges of remote interaction.

The Peerbots organization does not advocate for their tool to be used in any particular type of therapy or other assistive activities. Peerbots is used in a variety of therapeutic applications (such as art therapy or occupational therapy) and for a variety of different robot-interaction activities, such as social skills exercises, reading books, or role-playing interactions [11]. Peerbots is also used in both individual sessions with a single child or group activities with several children. Across these domains, Peerbots are used in a variety of ways; at present, there are also no official guidelines for how Peerbots users should introduce robots to children or discuss their teleoperated status—each user makes their own decisions about how to navigate these elements of child-robot interaction.

A key component of Peerbots’ use in real-world educational or clinical environments is that users play a significant role in designing the dialog-based activities their robot can be used for. Because individual users must adapt the Peerbots tool to their personal practices, they create their own *content*. In this case, content refers to the pallets of dialog buttons that are *authored* in the Peerbots teleoperator user interface. When users *author content*, they prepare all utterances that they anticipate making the robot verbalize during an interaction and create dialog buttons for each of these utterances.

The findings of our interviews and usability tests are relevant to the design of many types of assistive robots beyond Peerbots. However, grounding our work in this platform and user community allowed us to gain a more rigorous understanding of the routines and challenges that are necessary for assistive robot tools to be incorporated into real-world settings.

Table 2.1 Participants in interviews with early adopters of SARs: IDs are T for therapists, E for educators, and N for neither.

ID	Therapeutic Credentials	Experience with robots	Duration
Alissa (T1)	Registered Behavior Technician with 3-5 years of experience delivering behavioral and art therapy to neurodivergent children in one on one sessions.	Used robots in therapy with children in the past.	00:55, 9,899 words
Blanche (T2)	Dance and movement therapist with 5-10 years of experience working with children who have various disabilities.	Extensive experience with robots and regularly uses robots for group social skills sessions.	01:33, 19,132 words
Caroline (T3)	Therapist with B.S. in Psychology with 5-10 years working with children who have various disabilities.	Used robots in therapy one on one and in groups.	01:04, 8,346 words
Emily (T4)	Occupational therapist for 5-10 years with 3-5 years working with children with various disabilities.	Used robots for some one on one sessions.	00:36, 6,305 words
Fiona (T5)	Licensed marriage and family therapist with over 10 years of experience working with children who have various disabilities.	Regularly uses robots for therapy with children.	00:47, 7,418 words
Greg (T6)	Behavior interventionist with over 10 years of experience working with children who have various disabilities.	Led social skills group sessions for many years then left to found a social robotics company.	01:09, 8,931 words
Holly (E1)	Child development specialist with over 20 years of experience in teaching.	One year of using robots in one on one sessions with children who have various disabilities.	01:04, 8,456 words
Isaac (N1)	No therapeutic credentials. Adapted Aquatics Instructor	Early adopter working with children to delivered social skills content using robots to groups of children for 5-10 years.	00:58 8,133 words
Jaclyn (N2)	No therapeutic credentials.	Organization leader using robots to regularly deliver therapy for groups of children with various disabilities for 5-10 years.	00:58, 10,308 words

2.9 Interview Methods with Early Adopters

To understand the experiences of early adopters of socially assistive robots, we conducted ethics-board-approved semi-structured interviews with 9 participants who used teleoperated SARs with children in practice. All participants had experience or an accreditation in a therapy practice with children and had experience teleoperating robots during sessions with neurodivergent children. Details about our participants are described in Table 2.1. Of our nine participants, six were credentialed therapists, one was an educator, and two were non-therapists who have relevant experience in robot-assisted therapy for children.

Participants signed a consent form and data use policy, as well as completed a survey to report their background, therapeutic credentials, experience working with children, and experience working with robots. Our conversation covered conducting and preparing therapeutic social assistance with and without a robot. In the final portion of our interviews, we asked our participants about how they would use a specific teleoperation interface (Peerbots [115]), shown in Figure 2.2.

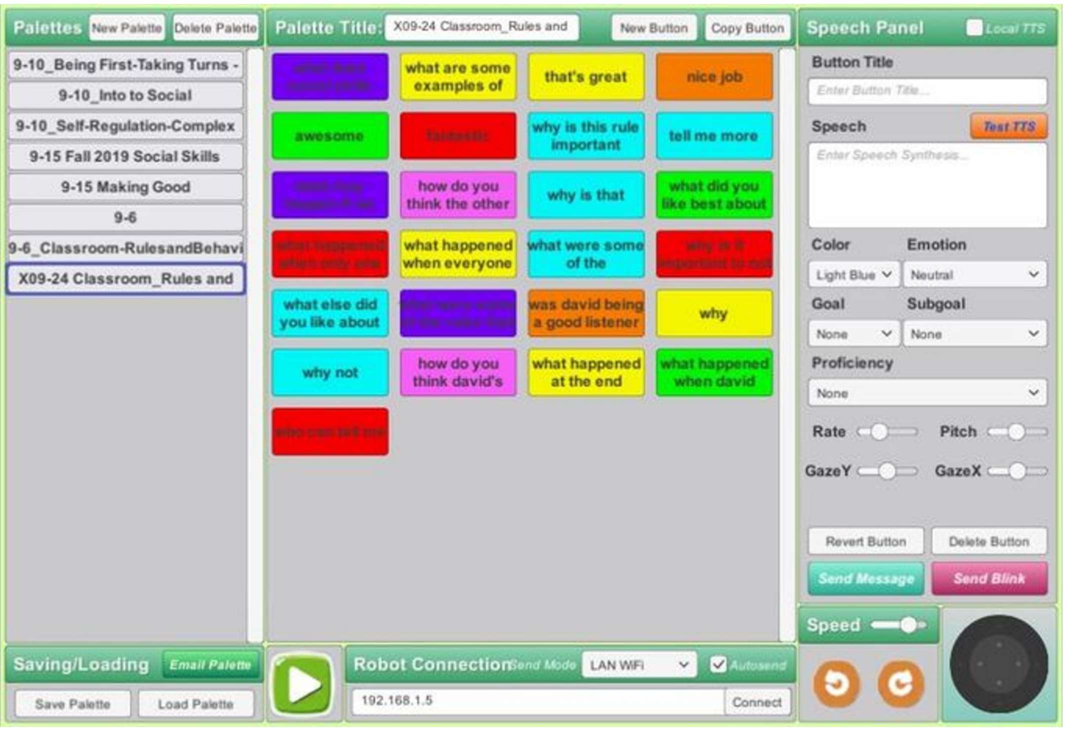


Figure 2.2 An image of the Peerbots teleoperation user interface for dialog control. On the left, various panels containing interaction content are shown, including “Classroom Rules” which is selected. The main section in the middle displays all of the dialog buttons included in that panel. An operator presses these buttons for the robot to verbalize their contents. On the right, another section of the interface includes a set of tools for creating new dialog buttons, as well as controlling the robot’s pitch, gaze, and movement (if available). The bottom of the interface displays connection information, showing the IP address of a separate device that serves as the robot’s face and voice.

2.9.1 Interview Analysis

We conducted and recorded interviews through a virtual meeting platform. Conversations ranged from 36 to 93 minutes, with an average of 60 minutes. Interviews were then transcribed and the transcriptions were anonymized. To analyze transcripts, we followed a grounded theory approach to identify key patterns from our participants' perspectives about therapy. We selected grounded theory since it is effective at working in a largely unknown context to produce an explanation of an underlying process [162]. We follow the general principles and heuristic devices of grounded theory [163, 164]. Researchers created open codes, which summarized individual observations from each sentence of interview transcripts. These open codes were collaboratively synthesized into axial codes, which represent consistent categories or topics. From these axial codes, we developed six themes, which we present in our results.

2.10 Group Usability Tests with Novice Users

The second component of our overall project was a pair of group usability tests with therapists who had never used a robot teleoperation interface before. We completed this inquiry to gather data from new users about the potential challenges they face when learning to use a teleoperated robot, as well as the ways they speculate about integrating robots into their practice. Group usability tests were conducted with 5 therapists who had no experience with robot-assisted therapy. All participants had experience as clinicians and at least one had experience as a teacher (anonymized participant IDs are shown in Table 2.2).

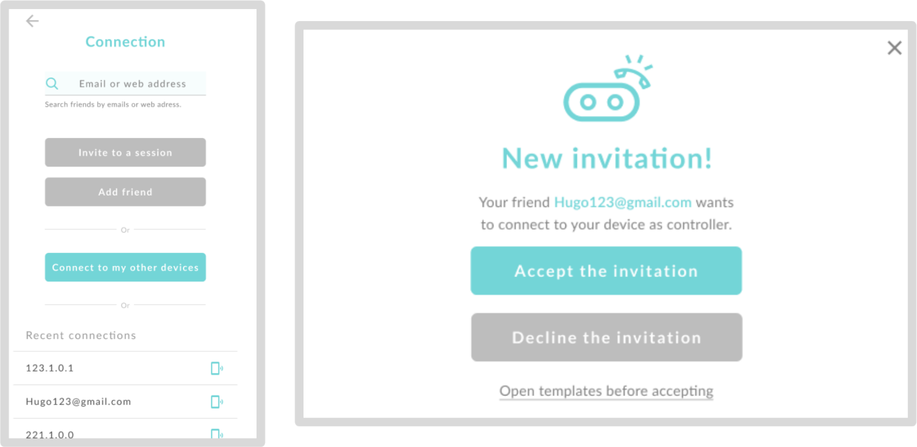
2.10.1 Group Usability Test Methods

We chose to use group usability tests [165] since they combine several benefits of focus groups and individual usability testing. The group setting encouraged discussion among participants. The usability focus provided an interactive setting for new users to develop a frame of reference for SARs and to explore how they might *act* when using a new teleoperation interface in non-hypothetical ways. We conducted our group usability tests through a virtual meeting platform. Participants signed a consent form at the beginning of the test. Each test took about 90 minutes and followed a similar structure, beginning with an explanation of robot teleoperation. Then, participants used an existing robot teleoperation platform to simulate the logistics of a telehealth interaction, in which the client and robot are co-located, but separate from the therapist (and their robot controller). We chose to use Peerbots, the same existing SAR tool as was used in the early adopter interviews [115]. The Peerbots teleoperation interface is open source and low cost, as it uses smartphones and tablets which may be readily available to users. Peerbots is already used in the wild for social skills programs with autistic children [11]. Following the simulated telehealth interaction, we presented participants with a variety of interactive Figma prototypes extending the functionality of the Peerbots application (Figure 2.3). The interactive prototype covered three common activities that novice SAR users must complete: 1) connecting to a teleoperation controller from a robot using email invitations, 2) connecting to a robot from a teleoperation controller, and 3) creating robot dialogue from the teleoperation interface. In this way, our tests allowed users to work with an existing high-fidelity tool *and* consider new potential features. Finally, participants and researchers discussed of their experience.

Table 2.2 List participants in group usability tests with therapists who held no previous SAR experience.

ID	Participant Name (anonymized)
P1	Emily
P2	Olivia
P3	Jennifer
P4	Chloe
P5	Jon

While logistics like robot setup and controller connection may seem unrelated to socially assistive interactions with children, they are critical to teleoperators' ability to integrate robots into their practice. Teleoperators' experiences with these logistical aspects of robotic technology affect their motivation to adopt it and use it over time. If users are unable to set up and connect to a robot, any improvements to dialog teleoperation will be inaccessible. Furthermore, the way teleoperators understand a robotic system, describe the system to clients, and adapt in-the-moment to technical difficulties are all critical parts of an interaction. Effective logistics, like setup and connection, ultimately support teleoperators' ability to create positive experiences for clients, be transparent about robotic tools, and adapt to inevitable technical difficulties.



Your friend Hugo123@gmail.com wants to connect to your device as controller.

Figure 2.3 Examples from the group usability tests of the invitation and connection screens in the custom Figma prototypes allowing users to connect to a robot an initiate a robot teleoperation session. © 2022 IEEE

2.10.2 Group Usability Test Analysis

Transcripts from the group usability tests were anonymized. As with the interview results, these transcripts were analyzed using a grounded theory method. Researchers created open codes that provided low-level summaries of individual observations. Then, researchers collaboratively synthesized these open codes into axial codes that represented broader ideas. From these axial codes, we developed four themes, which we present in our results.

2.11 Early Adopter Interview Results

2.11.1 Therapy as a Dual-Cycle Process

A critical overarching finding of our interviews with SAR early adopters was that understanding the *cyclic structure* of therapy is a prerequisite to understanding how robots fit within the existing practices of users. We developed a framework for understanding the cyclic nature of therapy as two nested cycles that take place over different time scales. Our analysis showed that this framework is key to the way therapists discuss their work and the role of technology within it. At a high level, therapy involves examining the client, evaluating their needs, preparing to meet those needs, delivering an intervention, and then repeating that process. This process occurs on a long timescale of months or years across the clients' sessions (the outer cycle), and a short time scale of minutes or hours within each session (the inner cycle).

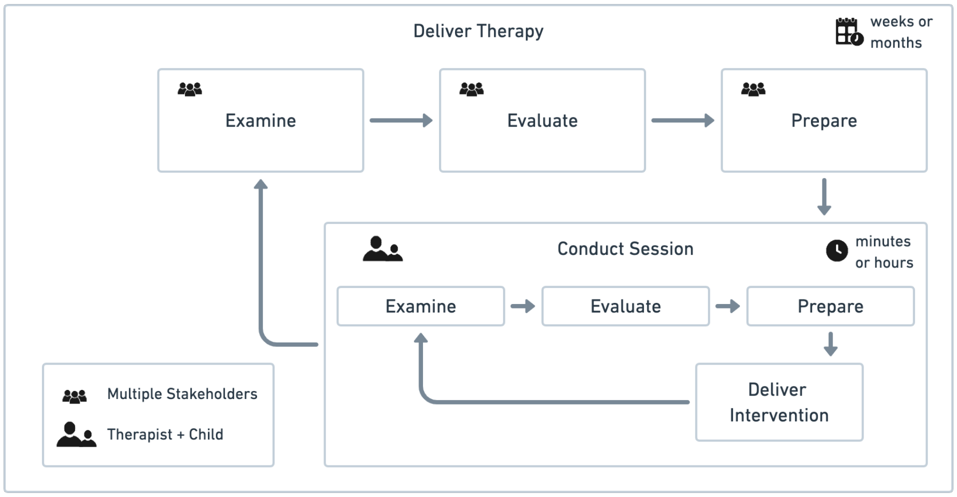


Figure 2.4 The cyclical process that therapists undergo in therapy with a client. The outer cycle involves a variety of stakeholders whereas the inner cycle within a particular session is focused on the therapist and their client.

The inner cycle involves only the therapist and child working together, whereas the outer cycle involves input from a wide variety of other stakeholders. A visual representation of these cycles appears in Figure 2.4. The outer cycle begins when a new client is identified. They are examined and evaluated by a specialist who can determine their needs in collaboration with parents, schools, and healthcare institutions. Afterward, therapists collaborate with these other stakeholders to outline the client's goals and determine the relevant interventions, deliver them, and evaluate their success. Within each session (the inner cycle), therapists undergo a similar cycle of examining their client's goals, preparing interventions, and delivering interventions.

Research on therapy is often approached with a focus on either the outer cycle of macro-interactions between stakeholders [166, 167] or the inner cycle of therapeutic interventions [146, 147, 147, 148]. However, these single-cycle perspectives miss a key piece of the story. By understanding how the outer and inner cycles interact from therapists' perspectives, we identify valuable considerations that can improve robotic technology used in this domain. Research that only focuses on child-robot interactions (within the inner cycle), fails to account for outer-cycle activities, such as preparation and evaluation, which are critical for the deployability of new therapeutic tools, especially robots. Moreover, a dual-cycle perspective emphasizes therapists, educators, parents, and caregivers as users of these systems in the outer cycle, not just children.

2.11.2 Interview Themes

The six themes (as visualized in Figure 2.5) that emerged from our grounded theory interview analysis are directly connected to the way that therapists examine, evaluate, and prepare in the inner and outer cycles of their practice. We will explore each theme in detail. Many of these themes reveal aspects of practitioners' experience in providing therapy that are non-obvious in a lab context that focuses on child-robot interaction, but that are deeply tied to the usability of robotic systems in the wild.

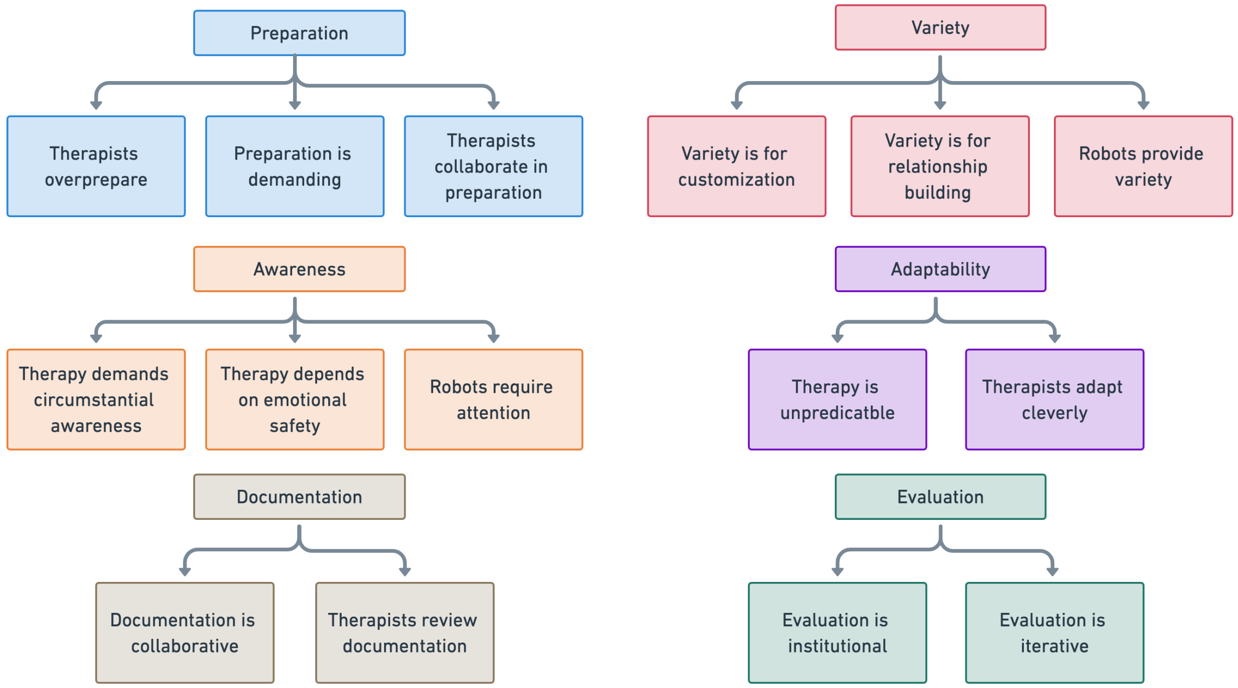

Preparation
Therapists overprepare
Preparation is demanding
Therapists collaborate in preparation
Variety
Variety is for customization
Variety is for relationship building
Robots provide variety
Awareness
Therapy demands circumstantial awareness
Therapy depends on emotional safety
Robots require attention
Adaptability
Therapy is unpredicatble
Therapists adapt cleverly
Documentation
Documentation is collaborative
Therapists review documentation
Evaluation
Evaluation is institutional
Evaluation is iterative

Figure 2.5 The themes and sub-themes derived from our analysis of interviews about delivering services using robots to neurodivergent children.

2.12 Theme: Preparation

Effective preparation is necessary for effective therapy. Yet, preparation is often omitted or disregarded in work focusing solely on individual interactions with children. In contrast, framing therapists as robot users naturally centers their preparation process. Our participants explained how demanding preparation can be, how the uncertainty and sensitivity of therapy lead them to over-prepare, and how introducing robots complicates their preparation process. When they use a robot, their preparation needs become even more necessary and time-consuming.

2.12.1 Preparation is Demanding

Preparation is a fundamental part of therapy; it isn't just a one-time occurrence but instead changes every week or session. Holly (E1) , a teacher with over 20 years of experience, mentioned that *"there's always preparation, no matter how experienced you are."* Preparing for therapy is time-consuming even without the incorporation of a robot; when a robot is introduced, it exacerbates this time burden. Holly (E1) explained: *"You know, I wasn't born and raised with technology like kids today. So, the technology end of it for me kind of felt like a lot of my focus in preparation".* Therapists are therapy experts, but not necessarily experts in operating robots or other technologies. As Blanche (T2) put it, *"I'm a therapist, like, I'm not like, like, I don't know tons of tech stuff, like, I am passable at best... I can troubleshoot small things, but when something massive is happening with an app that I know nothing about, I'm bewildered."* Developers often design robot therapy interfaces as expert-interfaces, that therapists need to master for the job. Even Isaac (N1) , who suggested being technologically savvy, mentioned that *"I took some practice because it was, it was a tedious process in the beginning."* Caroline (T3) shared that *"planning for that session (with a robot) for me was harder, really hard in the beginning".* Other participants reflected on how preparing robot content requires far more detail when compared to not using a robot. Alissa (T1) mentioned that with robots they had to *"pre-plan everything really extensively to be successful in those sessions."* Caroline (T3) , a therapist with over 5 years of experience, described how when using a robot, *"I actually have to plan even more."*

2.12.2 Therapists Over-Prepare

For therapists, being well-prepared means being over-prepared. When Holly (E1) described their preparation process with a robot, they emphasized: *"I typically over-prepare—I mean I have many activities in the bag... in case you pull out an activity and the child is not interested in the activity, you might try another."* Preparing a large variety of activities helps therapists connect with their clients (see Section 2.13.2). An important part of understanding therapists' needs around preparation is understanding just *how much* they over-prepare. Blanche (T2) explained that *"if the session is going to be 45 minutes or an hour I'm going to have probably 85 minutes worth of music...I always over-prep everything. I'd rather be over-prepared than under-prepared."*. That's close to double the amount of content used in-session. While having more content can help therapists feel prepared with a robot as well, it makes it more difficult to organize and access content easily in sessions. Blanche (T2) revealed that *"I would write so much (session content) But then, sometimes in the moment it's hard to scroll through screens to find what it is that you want."* Some teleoperators, such as Jaclyn (N2) , organized all content chronologically, in a single collection. Some, as Blanche (T2) mentioned, created activity-based collections that were modular. Others chose to create modular collections according to content theme, such as content for feedback, redirection, or relationship-building. Although therapists' over-preparation is time-consuming, it nevertheless provides them with the tools necessary to form a strong relationship with their clients (see Section 2.13.2) in order to create an emotionally safe environment (see Section 2.14.2).

2.12.3 Preparation is Collaborative

Finally, the therapy preparation process is collaborative. Our analysis shows that collaborative preparation is vital to therapy and occurs during the outer cycle. Our participants described preparation as a matter of iterative input from many stakeholders. Goal setting, which is a primary component of preparation, can rely on parents, teachers, insurance companies, and others. For example, US children's special education teachers organize their goals into Individual Education Plans (IEPs). Holly (E1) described how they begin with *"knowing what (the client's) IEP might say, having anecdotes from teachers—that's preparation."* Therapists often rely on this crucial documentation from others (see Section 2.16.2). Parents often provide substantial input into their children's therapy. Greg (T6) told us how *"the parents can tell you what worked for (the client) already."* While parents and therapists have direct interactions with children over the course of therapy, other indirect, yet significant stakeholders are also involved, such as health insurance companies. In the USA, where all of our participants practice, insurance approval is often required for children to receive therapy, as well for the use of a robot in therapy. Participants described coordinating with insurance companies to approve a robot as a therapy tool or to incorporate it into a treatment plan. Greg (T6) connected the approval process back to goal-setting: *"Most therapy is paid for by insurance or by the state and so it's fairly structured. You have to be working towards very specific goals and those goals have to be agreed upon by the team"*. The preparation and goal-setting process, which involves both institutional and familial input, is additionally tied to whether a client receives funding for their therapy.

2.13 Theme: Variety

All of our participants stressed that having a variety of activities and interventions for a session prepares them to proceed smoothly and work effectively. Therapists emphasize that variety is critical in order to customize sessions and build relationships with their clients. Robots provide access to a variety of content which makes the robots appeal to therapists for use in therapy.

2.13.1 Variety is for Customization

Therapists choose different interventions for different children based on factors such as children's age, the challenges they face, their therapy goals, and the features of their IEP or insurance plan. Emily (T4) reflected that the challenges a child faces *"present in so many different ways; and then what's important to one child might not be important to another"*. All interviewed therapists emphasized that there is no single criterion by which to categorize children and determine whether they would benefit from specific types of activities. Participants were reluctant to make generalizations based on children's age or disability; some even challenged interview questions that implied such generalizations. Greg (T6) pointed out that *"with autism, there's such variety. You can have a ten-year-old working on the same skills as a five-year-old, it's really dependent on the child."* Alissa (T1) shared that *"it's important to not force kids into a box, that we want them to fit in. Like I think it's important to let a kid be who they are, and to meet them where they're at in any type of therapy that they get."* Technologists ought to be careful when deciding to structure (or market) products according to disability or age. Therapists themselves rarely rely on such generalizations, which are often unhelpful or inaccurate for individual children.

2.13.2 Variety is for Relationship-Building

Content and activity variety provides the structure to build a strong child-therapist bond. Greg (T6) brought up similar points about the relationship between variety and trust: *"in* *between all the structured exercises, you want to maintain the relationship with the child, so you have to find ways to make it fun and build that trust relationship. So having a lot of variety is always good"*. Our participants describe therapy as an experimental process where therapists try new activities and see what engages their client. As Greg (T6) put it, therapists are *"just constantly trying new things until you figure out what works and what turns (clients') mind on"*. Though this need for variety contributes to the demands of preparation, offering variety fosters trust and promotes a positive client-therapist relationship.

2.13.3 Robots Provide Variety

Robots' ability to offer a large variety of activities contributes to their appeal for use in therapy. Our participants reported that they use robots to facilitate games, read books with children, listen to children read books, start conversations, and conduct many other activities. Fiona (T5) mentioned that *"there's a lot of different options with this robot and software; everything from academics, to social skills, to vocational skills, self, daily living skills, stories"*. However, with robots, each individual activity is rigid in terms of customization. Emily (T4) described without a robot, customization is simpler *"since I'm the one verbalizing everything I can customize it because I'm verbalizing"*. Holly (E1) described their difficulty trying to customize activities for children using robots by explaining that *"with responses of the robot you only have one or two keys, you can press"*. These difficulties can slow down therapy as therapists try to figure out how to use the available tools.

2.14 Theme: Awareness

Therapists are highly alert to everything going on during a session. They must monitor their client's progress, emotional state, keep track of the features of the teleoperation interface, and resolve any technical difficulties. Additionally, they must maintain the pace of conversation and avoid missing important or emotionally sensitive moments in a session.

2.14.1 Therapy Demands Circumstantial Awareness

When we asked Alissa (T1) about what they need to be aware of in sessions they said *"Everything. Everything at the same time"*. When we asked Greg (T6) about the sorts of things they are aware of during a session, they told us *"Everything! You have to be aware of everything* *in the environment at all times. You can be missing 90% of the picture if you're not tuned in."* Greg (T6) added that they also learned to be aware of contextual factors in a child's life (*"what their day has been like, what their diet has been like, have they eaten at all today, do they need to go the bathroom, did they just have a fight with their brother, sister. . . "*) One way therapists do that is through *"anecdotal remarks from other staff, like So-and-so just got moved into a new foster home this week"'* says Holly (E1) . All these factors inevitably add to the cognitive load that therapists experience in the inner cycle of therapy. It is crucial for technological therapy tools to help reduce, or at least not substantially increase, demands on therapists' awareness.

2.14.2 Therapy Depends on Emotional Safety

Emotional safety precedes all other activities in therapy. Fiona (T5) described how most of their awareness goes towards "reading" their clients, *"So, you always want to read them . . . because you want to build rapport. If you don't, if the child doesn't feel comfortable, then everything else you're trying is not going to be effective."* This form of deeply empathetic awareness builds the foundation for therapy. Without it, clients cannot work toward any of their goals. 4 of 6 therapists described how their relationship with clients makes therapy both possible and meaningful. Blanche (T2) affirmed how *"the most important aspect of therapy for anybody is that experience of being seen, being witnessed—being allowed to be your whole self."*

Children may sometimes feel emotionally safer with robots than with humans. Caroline (T3) said that *"The way the robot presents itself doesn't come with like this, many years of social representations you know. It feels safer for some clients."* Caroline (T3) continued that *"as any human, we do have some type of social expectations. And some children really struggled in responding to those social expectations, even if I tried my best."* When robots are used in therapy, they change the client-therapist relationship. Emily (T4) described the change as a *"perceived power shift. So, the robot kind of creates this new dynamic that we would never be able to achieve in the session, because I'm bigger, and I'm older and they know that, so it helps. The robot is less intimidating, not constantly watching every move the child makes, and so that helps"*. Blanche (T2) mentioned that *"The safety that (clients) feel with a non- human entity, you know, they there's a level of automatic trust that I saw a lot of our kids have, that was safe for them"*. Jaclyn (N2) explained robots' impact saying, *"the robot was able to open them up; reduce their* *inhibition and motivate them to verbalize"*. This is consistent with the research [73] and is a large factor in why therapists use robots.

2.14.3 Robots Require Attention

When teleoperated robots are used in sessions, they place further demands on a therapist's attention during an already stressful inner cycle. Emily (T4) mentioned that *"the robot is just kind of like an added step, so I actually think it adds to what I need to be aware of"*, then later said *"but I don't think that's a hindrance"*. This sentiment that robots are difficult to use but are worth it was common. Robots may be considered valuable because of how they change power dynamics with children, or due to the increase in variety of options that robots provide therapists (see Section 2.13.3) to connect with clients. However, robots introduce additional logistical concerns for therapists. Since therapists aim to have robots match the client's pace of conversation, teleoperation interfaces themselves inherently demand attention. Emily (T4) told us about this challenge: *"That to me is the biggest struggle as a therapist—to kind of maintain that that timeliness of response in the conversation when you're trying to like type out a new button."* Greg (T6) had similar reflections about their use of robots; *"There's definitely a tempo that you have to keep to keep engagement going...If you're fumbling with the interface and you can't figure out how to say the thing that you want to say, but there's a moment that's happening right in front of you, and you want to address it and you want to make it teachable."*

2.15 Theme: Adaptability

Therapists describe therapy as an unpredictable, dynamic interaction in which maintaining a positive relationship is key. As the client's needs change within a session, therapists have to adapt and respond in the moment. With robots, therapists have developed clever, narrative strategies to make up for technical difficulties and lags in child-robot conversation.

2.15.1 Therapy is Unpredictable

In a session, a therapist's success at delivering therapy heavily relies on how they respond in the moment. Blanche (T2) said that *"it's cute to sit there and, like, for me, before session to type in all these different possible responses and I do. But then when I get there in the moment that doesn't mean that I'm going to have the right response."* Children may also test the therapist's knowledge and preparation as Isaac (N1) illustrated, *"I learned the hard way that the kids will test the robot and see if it does actually know everything"*. Jaclyn (N2) also mentioned the importance of being flexible within a session *"we kind of have to be flexible and deal with whatever we get but, but we do try to take everything into account that we possibly can to make things appropriate."* Therapists have to respond to whatever is happening in the session as illustrated by the inner cycle (examine, evaluate, then respond). When asked if this is stressful, Fiona (T5) alluded to a need for easier-to-use tools: *"I don't know, maybe there's a quicker way to even find it here I'm not sure."* Holly (E1) mentioned that *"most kids are pretty forgiving, so I don't think you have to be perfect at it"*. This likely explains why current tools continue to be used despite their imperfections. The benefits of children interacting with robots can outweigh the issues that therapists have to deal with.

2.15.2 Therapists Adapt Cleverly

Therapists adapt their activities according to what they notice in the moment. Blanche (T2) mentioned that *"You go in there with like a general goal, maybe for the student, but you have to be spontaneous and you have to be ready to respond to what's there"*. The use of robots introduces another facet of therapy that therapists have to adapt to, especially when technical difficulties arise. Our participants talked about how they manage technical difficulties, low batteries, or teleoperation delays. Holly (E1) also described that when problems occur they *"certainly can impair the whole process if you're if there's a glitch in the technology"*.

With certain technical complications, therapists are able to continue to use the technology to explain away the problems through a narrative. Alissa (T1) said: *"I use the power nap one a lot. Like right before I knew (the robot's) eyes would die, because I could see like the 10% in the corner, I would have a button that says "oh I'm getting so sleepy! I need to take a quick nap!""* Emily (T4) mentioned utterances they authored specifically to compensate for delays: *"I did have some buttons that said, like "Hmmmm... let me think about that", like after the child asks a question. I was doing a lot of free typing in and adding buttons on the fly kind of which definitely got a little stressful."* When therapists teleoperate robots, pacing is critical to providing a natural-feeling, positive experience for the child.

2.16 Theme: Documentation

Documentation is a crucial component of the planning process. Documentation is necessary to justify merit or validity of a particular therapeutic intervention, such as robots, to health insurance companies. In this way, documentation of the use of SARs is essential for clinics and other institutions to have financial access to such tools. Furthermore, documentation is necessary because therapists assigned to a particular child may change over time. Good documentation allows a new therapist to easily pick up where the last therapist left off. As developers, we emphasize the difference between documentation (writing down what happened) and evaluation (assigning a valence to that outcome). When technology is used in therapy, tools may support therapist-authored documentation. However, evaluation should be performed by the therapist themselves.

2.16.1 Therapists Review Documentation Before Sessions

Documentation, like preparation, is an essential part of how therapists stay organized. This is especially the case because therapists work with many different clients. Blanche (T2) shared that *"I usually review my progress notes from the session before just so I can catch myself up, because details might escape you a week between sessions"*. This review is to *"help me assess baseline levels for the top of the session"*. Greg (T6) pointed out that documentation happens both before and after each session, *"For two hour sessions, usually hour and a half doing direct therapy and 30 minutes documentation. 15 at the beginning 15 at the end."* Alissa (T1) mentioned that they *"run trials on different activities and things of that nature and I collect data on that"* implying that this data collection occurs during the session. It's clear that therapists have notes and collect data, because they're reviewing those at the beginning of the session, but isn't clear from our interviews if there are explicit standard ways in which they do so. Regardless of the format in which they collect this information, we know that it is crucial for them to review documentation, which plays a role in preparation being demanding (see Section 2.12.1).

2.16.2 Documentation is Collaborative

Our participants described several different ways that their teams collaborated to document therapy. Holly (E1) described how educators are an important source of documentation: *"If* *you're the head teacher or a teacher's aide or you're in the classroom every day with that child then your homework would be that there's documentation. Okay, and everyone should be aware of the documentation, whether the child is you know, has atypical or typical learning"* Fiona (T5) had a similar experience asking others about their clients, *"I'll also talk to teachers, how is so-and-so doing in this behavior, how they are in the class, as well as from administrators, because we all have different relationships with them, and we see different sides of them".* The documentation process serves to synthesize knowledge and observations from multiple therapists, educators, parents, and other stakeholders in each child's life.

2.17 Theme: Evaluation

Evaluation is a collaborative and iterative process that ties together therapists' responsibilities within and between each session. Institutions provide support with evaluation by creating standard procedures to ease therapists' evaluation and maintain consistent records about clients. With iterative evaluation, therapists are able to monitor their clients' progress, and report to other relevant stakeholders.

2.17.1 Evaluation is Institutional

Institutions can provide support for therapists and maintain a consistent experience for children over the course of their therapy. Institutions can also provide a standard format for evaluating clients to ease the burden on therapists. Jaclyn (N2) described *"summary forms —those are evaluation forms and those we have for individuals and we have for groups, and we have those for any kind of robot engagement that we do".* While these per-session evaluation forms are important, being able to analyze the success of therapy over a long period of time is also crucial. Jaclyn (N2) described *"evaluations after each program"* where programs are six, eight, or ten weeks. As the director of an institution, Jaclyn (N2) mentioned running *"a pilot for our multi-sensory session, that was a six-month pilot but we broke it into like two three-months summers for that."* If the evaluation shows a successful delivery of the therapy, then these programs are then deployed.

Institutions provide continuity when working with the client and other stakeholders. When discussing preparation for a new client, Greg (T6) mentioned that *"you're usually reading the* *notes from a therapist that worked with the child last. So you know if they picked up on something definitely want to try to build off of that."* The continuity across therapeutic evaluations is helpful especially when a child's therapist may change. Alissa (T1) described their experience entering an organization that had been working with robots: *"I kind of came in. Maybe in the middle of when they had started, but there were some schools that had already seen (the robot) for like a year and a half or two years before I started. So when I came in, they already kind of had this like established relationship with this robot."* When describing the experimental process to connect with a child, Greg (T6) even mentioned a last resort option to *"swap out therapists"* for the child. While ideally, children would work with the same therapist over time, institutions provide an ability for them to work with other therapists as well.

2.17.2 Evaluation is Iterative

Therapists report that it is crucial to continuously evaluate clients to monitor their improvement and ensure that they are receiving appropriate interventions. Greg (T6) mentioned that a child's overall *"goals are determined periodically throughout the year"* and that they are *"very specific to the children"* resulting in *"a lot of variety"* of goals. Therapists described two key ways in which the results of their assessments are used; insurance and preparation. Alissa (T1) explained that *"insurance requires an updated treatment plan every like four to six months"*. Alissa (T1) shared that an update to a child's treatment plan will be made up of *"a whole new set of goals that are put into the treatment plan that he needs to work on, because he made progress."* Insurance needs these updates to ensure that the funding they are distributing to the child is worthwhile and the therapy is effective. Therapists also want to make sure they are delivering effective therapy and use assessments to determine upcoming interventions.

2.18 Interview Results Summary

In this section, we have reviewed six key themes (noted inFigure 2.5) that emerged from a grounded theory analysis of interviews with early adopters of SARs with children. These themes illustrate the challenges faced by users who are already integrating teleoperated robots into their practice. In particular, our findings emphasize that many of the most labor-intensive parts of robot use occur outside of individual interactions with children—such as preparation, documentation, and evaluation. Furthermore, our findings emphasize that therapists require the ability to maintain awareness and adapt quickly to develop continuity and emotional safety during therapy.

These interview results highlight the experiences of therapists and other practitioners who already have experience creating content for robot-based interventions and teleoperating robots during interactions with clients. However, it is also important to consider the particular challenges that new users may face in this domain. To this end, we present the results of our group usability tests involving therapists with no previous robot experience (our methods are fully described in Section 2.10.1 and participants listed in Table Table 2.2).

2.19 Group Usability Test Results

Grounded theory analysis of the results of our group usability test revealed four key themes (shown in Figure 2.6). These themes illustrate many of the potential challenges that therapists or other practitioners may face in incorporating SARs into their work. As with the results of our early adopter interviews, many of these themes also emphasize aspects of the therapy process that exist outside of individual child-robot interactions. Our findings emphasize that robots are naturally compelling, but that the preparation and setup process to use them can cause friction. Similarly, therapists expressed concern about how they would compile and organize a sufficient variety of robot teleoperation content, as well as how they would coordinate and collaborate with other stakeholders in the therapy process while using a robot.

2.19.1 Theme: Novice Users Intuit Robots' Appeal to Children

Throughout the group usability tests, it was clear that participants understood the benefits that robots could provide. Emily (P1) mentioned that they *"imagined that we would use it for social skills"* or that the robot could be used in speech and language pathology to *"kind of help with some speech sounds"*. After using the system for a while longer, Emily (P1) added *"I would also probably use this to support like emotion recognition, emotion identification"*. Jennifer (P3) echoed these suggestions, saying they would use the robot for *" a virtual therapy session or virtual group therapy, like a social skills group or for a kiddo who's non-verbal."* Later in the discussion, Chloe (P4) asked *"What about, like, the robot reading a story to the child?"*

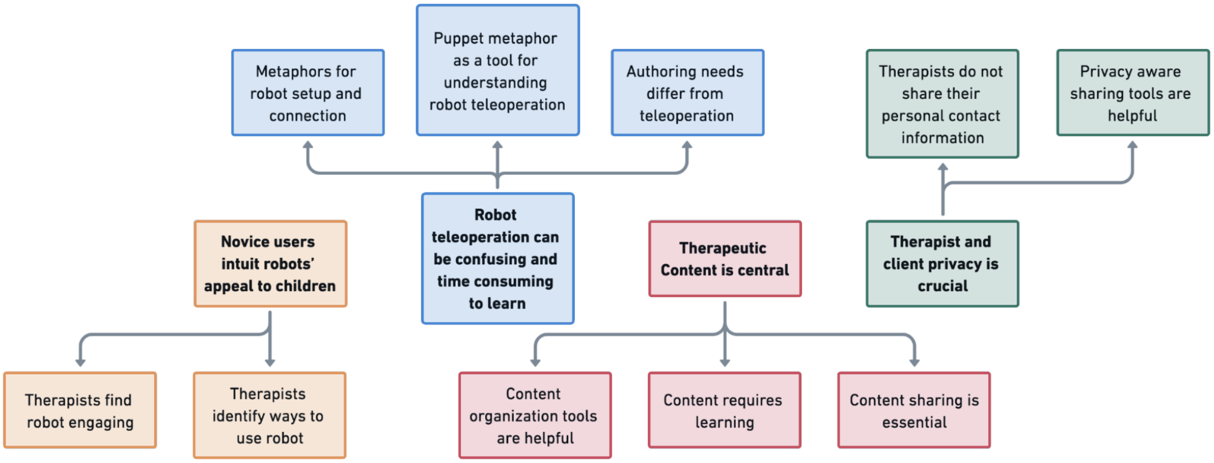

Metaphors for robot setup and connection
Puppet metaphor as a tool for understanding robot teleoperation
Authoring needs differ from teleoperation
Therapists do not share their personal contact information
Privacy aware sharing tools are helpful
Novice users intuit robots' appeal to children
Robot teleoperation can be confusing and time consuming to learn
Therapeutic Content is central
Therapist and client privacy is crucial
Therapists find robot engaging
Therapists identify ways to use robot
Content organization tools are helpful
Content requires learning
Content sharing is essential

Figure 2.6 A summary of themes we identified following our grounded theory analysis of group usability tests with therapists who held no previous SAR experience. © 2022 IEEE

Therapists know that engagement is critical in therapy [132, 168]. Our participants shared that it becomes even more critical when providing remote services. Emily (P1) described, *"especially if you're doing virtual sessions... it's very hard to work with young kids through a zoom screen."* Emily (P1) then went on to say that *"this could be a really helpful tool in that scenario."* These results show that technologists do not need to *convince* therapists of SARs' efficacy. It was clear that therapists had ideas about how they could incorporate a robot into their usual sessions and use it to engage with children. Instead, technologists should consider focusing on to making systems easy for therapists to use effectively.

2.19.2 Theme: Teleoperation is Confusing and Time Consuming to Learn

Therapists understood the benefits a robot might provide; however, they also understood the substantial difficulty of incorporating a robot into their typical practice. Participants described the robot setup process as particularly difficult. Jennifer (P3) said that *"walking through this stuff definitely makes sense."* but later added that *"I don't think I'd understand how to like, get the technology that I need to actually implement it today"*. Having a guided user experience may work once, but if users are to adopt SARs into their daily practice, SAR setup and teleoperation must be independently accessible to users. Setting up a robot is a difficult process for new users. In our group usability tests, we evaluated three different potential processes for making the setup process more intuitive, each based on a different metaphor: using a puppet, joining a virtual room, and sending an invitation. While these metaphors were helpful in some cases, we observed that the technical procedures associated with robot teleoperation were opaque and intimidating, especially when they relied on technical jargon that may be commonplace to computer scientists, but not to therapists.

During the group usability tests, we used the metaphor of a puppet to describe how the robot's body and teleoperation interface worked together. With low-cost alternative robot solutions such as Peerbots [115], users can use already available hardware such as tablets or phones and place them in a hollow doll body, hence the similarity to a puppet. However, this sometimes created confusion regarding the robot's autonomy. After Chloe (P4) connected to Jon (P5) 's robot and initiated a verbalization, Jon (P5) responded cheerily to the robot, seeming to perceive it as autonomous. After a demo of the setup and after connecting to another participant's tablet, Jon (P5) was still confused about the robot's teleoperated nature saying *"I guess I'm still. I'm still confused about the, the backpack robot. How does, so does that, does that eventually show this face? and does that talk?"*. New users may need time to understand robot teleoperation, especially when the robots' body and dialog controller are different devices. Similarly, the idea of joining a virtual room was an unintuitive metaphor for the robot setup process. This process required additional information from other participants, especially in the telehealth scenario. For example, the Peerbots application generates a random room code so that users do not accidentally enter rooms that are already in use. However, participants found it troublesome and annoying to share random combinations of letters and numbers for room names. Participants shared room information multiple times due to difficulty clearly hearing the generated room codes. Jennifer (P3) went through three attempts of sharing their room code, *"Okay. So let me try it one more time. All right. New code is W one M B as in boy, P five."*

We also considered the metaphor of sending connection invitations, an approach in which users could enter an email address to send an invitation to that address. This went far more smoothly with most of our participants. All (5) participants found it easy to receive an invite and only one had trouble sending connection invites. While the invitation method was easier to follow for our novice users, one participant expressed some concerns. For example, Jon (P5) described a button to "connect to all other devices" with concern, saying it would lead to *"Bad news. I don't like clicking on anything that says connect to my other devices. So what does that mean? and* *what devices? and all that. I have a lot of cognitive dissonance right now, I've got to tell you."*

Our group usability tests produced two additional connection metaphors: calls and web URLs. Since invitations could be declined, participants wanted the ability to respond with a new invitation, as well as make sure that there was a log of all past invitations; similar to a phone log of received and missed calls. Another metaphor that made sense to our participants arose from how we shared the interactive Figma prototype used for our group usability tests. Participants clicked a web URL and the prototype would directly load with no additional setup. Jennifer (P3) shared that *"If it was the computer thing where I could just like send the link to the family, then that'd be great."* Overall, technologists should frame the robot setup process intuitively for users in different use contexts. For example, Emily (P1) mentioned that it would be important to make sure *"parents are trained in how to use it"* for virtual interactions.

2.19.3 Theme: Therapeutic Content is Central

When participants first interacted with the teleoperation interface, they said the options were *"overwhelming"* or that they found the interface *"more complex"* when compared to the robot face. Jon (P5) described the dialogue teleoperation buttons as a *"mass confusion"* because they *"have no idea what it is"*. It became clear that learning the interface was a separate challenge from learning the therapeutic content itself. In addition to understanding the operation of the dialog control interface, therapists still needed to learn what therapeutic content was included in the interface for them to deliver.

As teleoperators, therapists expected content to already be present in the application. Chloe (P4) asked, *"is there like a template, a lesson sort of, that's already set up that we could use? sort of just like, social comments that I might say"*. It also became clear that participants wanted a variety of content. Emily (P1) mentioned that *"it would be helpful to have some more differentiated collections"*. Therapists also discussed how content-organization tools would be essential to save time. Chloe (P4) described how *"we don't have a ton of prep time, right, for sessions. So things need to be like, ready to go."* Olivia (P2) followed up with *"we don't really want to spend much time preparing . . . the biggest reason for me like to choose or not to use a specific app is like really, how could it help me save time and work more efficiently?"* And the clear answer to helping with time was, *"some resources that I can use or template or whatever* *things that could facilitate my work".* While the teleoperation interface was necessary to use the robot, it was clearly not sufficient. Participants anticipated relying heavily on the *content* within the teleoperation interface.

Sharing intervention content and materials is common practice in therapy. Jennifer (P3) described how *"especially for like, a social skills group, or something like that. If you create your own lesson plan for the day or for the week, it's very helpful to give that to someone else who might be doing a similar group."* Because therapists often adapt one another's content, it is important for robot teleoperation interfaces to, *"have an editable option where I can tweak it to better fit my needs".* Chloe (P4) described having some starter content that *"feels much better than starting from scratch."* When discussing feedback about the teleoperation interface, Chloe (P4) interrupted our discussion with an urgent suggestion, *"sorry, just one more thing, if you could make it shareable. Like, if I created a book that I thought was a great resource, that's working well for me, if I could share that . . . so other people could grab it."* to which Olivia (P2) responds, *"Yeah, that'd be awesome"* with some excitement. Sharing robot dialog content was clearly necessary for therapists.

In addition to the ability to share content, participants also saw personalized content organization as an important part of robot usability. Jon (P5) described that *"if there could be a heading, like a heading that you could organize the rows or the columns, I would like that".* Color coding different dialogue was another approach that seemed useful. Chloe (P4) explained, *"if I'm trying to quickly keep up with a back and forth exchange with the child, different color codes would help me. Like if I grouped like reds for all like the social exchanges. . . ".* But without a legend, looking at others' authored content Chloe (P4) and Jon (P5) were left wondering what each color represented. To organize their content, Olivia (P2) wanted a *"few folders . . . where I can categorize the things together".* Participants suggested different approaches to organization. When describing their approach, participants noted two key factors that influenced how they wanted content organized. Emily (P1) described that *"It might depend on the lesson".* For example, for a *"social interaction"* they might *"put them together in a scripted manner"* but if the content was *"a more broad panel of buttons"* then they would *"organize by type".* When asked about how they would want content authored by others organized, both Emily (P1) and Jennifer (P3) agreed that they would want content *"grouped by category".* When learning content from other authors, participants also wanted the ability to quickly find dialogue: Emily (P1) also followed up by asking if *"you can do command F and search for certain words?"*.

Teleoperation through the interface seemed intuitive to participants once the robot was connected. Users automatically navigated different content collections and selected buttons for dialogue they wanted to try. However, it was not clear how such content was created. When asked about the content editing capabilities of the interface, Jon (P5) mentioned that they *"don't like that. I don't want to edit anything that I don't know what's going on. I get very scared."* Some participants expected teleoperation to be completely separate from authoring and that they would only use pre-made content. After teleoperating the robot for a short period, other participants intuited the need to author custom content. Emily (P1) walked us through how they would author content step by step without needing any assistance. After some time exploring the interface, they also described how they would use different authoring features; *"I would use this [describes edit panel] because I'm kind of like creating a new thing. If I were in the middle of a session, and like needed to make a new response, I'd probably gravitate towards [refers to open-ended dialogue entry feature]"*. Participants intuited that fast authoring during a session and more thorough authoring in preparation for a session are distinct needs.

2.19.4 Theme: Therapist and Client Privacy are Crucial

Participants identified privacy, both the therapist's and the client's, as another key factor in designing robots that can integrate with the norms of therapy practices. Therapists require control of the information provided to clients, to protect their own privacy. Similarly, therapists must also store and share their clients' personally identifiable information carefully. Participants described communicating in a variety of ways with their clients and clients' guardians to plan and coordinate sessions. Chloe (P4) shared that they *"give parents my business card with my email and my phone number, my work phone number. But I do not give them my personal cell phone or my personal email address."* This emphasis on creating a personal boundary with parents seemed important to some participants. Specifically, Jennifer (P3) mentioned that they *"don't give my email out to families because it can be misused sometimes"*. They even followed up saying that they would create a separate email for using the robot application if the application required email. When asked, Jennifer (P3) mentioned that having usernames instead of emails would be beneficial. Aside from therapist-imposed boundaries, some therapists seem to have restrictions on interacting with clients that are beyond their own control. Jon (P5) mentioned that *"we're highly restricted from any communication. So really I can't do email. . . . I can't do anything now, but just talk to people"*. Others mentioned providing their organizations' email address or contact information as well.

Clients' information is also crucial to keep private. Participants emphasized how they are highly aware of clients' personally identifiable information and are cautious about storing and sharing it. While sharing content is a crucial component of preparation for therapy, therapists require easy ways to remove content that may contain private information or use other methods to deliver client-specific content. Specifically, Emily (P1) mentioned that they would *"probably just delete [a private] button before I shared it with anyone."* Jennifer (P3) followed up by saying *"If I was going to say something that had some privacy information in it, I would honestly probably do the [open-ended dialogue entry feature], as long as it didn't save."* When we suggested having some buttons designated as private and would not be shared, Emily (P1) shared, *"I probably would use it just to have like some token, like hi so-and-so, or great job so-and-so so I didn't have to type those in every time"*. Overall, participants in our group usability test described a variety of privacy concerns, on both their own and their clients' behalf, that could impede their ability to use social robots in practice.

2.20 Discussion

The goal of this project was to explore the use of teleoperated social robots in assistive domains with children. Through interviews with early adopters and usability tests with novice users, we highlight the values and needs of therapists, educators, and other caregivers. Our findings affirm that centering these adult practitioners as the primary users of socially assistive robots is necessary to understand how robots can successfully be developed and deployed in this emerging domain. Therapists find robots engaging and valuable as tools that can engage children's attention and facilitate therapy. However, they also face numerous challenges in adopting robots into their practices to access the benefits of assistive child-robot interaction as demonstrated in laboratory research. Ultimately, roboticists must meet the needs of adult robot teleoperators for SARs to successfully assist and support children in therapeutic settings.

A key component of our findings is that describing therapy as a dual-cycle framework reveals how robots may align or cause friction within the personal and institutional practices therapists follow. Our dual-cycle framework is composed of two nested cycles each taking place over different time scales. Therapy involves iterative examinations of a client, evaluations of their needs, preparation to meet those needs, delivery of an intervention, and evaluation of its success. This process occurs both long-term across therapy sessions (the outer cycle)and short-term within them (the inner cycle). Robot teleoperators have different needs and face different challenges at different points in these cycles. They require robotic tools that are compatible with the different types of complex, demanding tasks that take place during these two phases of therapy. When technologists consider tools for robot-assisted therapy, they often focus on the inner cycle of sessions without consideration for the outer cycle and how it influences therapists' preparation, documentation, evaluation, and relationship to other institutions. However, technical features relating to the preparation of session content or documentation and evaluation of clients' progress are integral to the success of therapy tools.

2.21 Therapists Must Work Within Collaborative and Institutional Processes in the Outer Cycle

The outer cycle of therapy is characterized by long-term collaborations among a diverse set of stakeholders including parents or guardians, teachers, therapy institutions, and insurance companies, each with different perspectives about the child. This is the case during the preparation (Section 2.12.3), documentation (Section 2.16.2), and evaluation (Section 2.17.1) phases of therapy. Through preparation in the outer cycle, therapists adapt sessions to their client's changing long-term needs. As part of the preparation for therapy, therapists have to review documentation from previous sessions or other collaborators (Section 2.16.1). This review increases the demanding nature of preparation (Section 2.12.1) since the client's evaluation throughout therapy is iterative (Section 2.17.2).

Both familiar and novice users felt that preparing a sufficient amount of varied content for robotic therapy tools was a challenge. In interviews, early adopters shared the demanding nature of preparing to use a robot with children. Similarly, novice users emphasized their concerns about having access to quality, well-organized content when learning to use a robot. In addition to requiring a wide array of robot content, therapists must coordinate their preparation process with objectives set collaboratively by other stakeholders, as well as with any requirements from insurance companies that may determine whether robotic interventions are financially feasible. In this way, considering teleoperators' preparation needs outside of individual interactions with children demonstrates how robotic platforms must support users in long-term, outer cycle activities.

Similarly, robots used in therapeutic contexts must be compatible with the documentation and evaluation needs that occur in the outer cycle. Documentation and evaluation procedures are a critical component of therapy with SARs—they allow therapists to collaborate (Section 2.16.2), to measure and report progress to other stakeholders (Section 2.16.2), and to maintain insurance approval for robotic tools (Section 2.17.1). Early adopters described the role that documentation and evaluation play in planning appropriate activities and in making well-informed decisions in collaboration with a child's educators and guardians. Novice users emphasized concerns that robotic tools must also allow therapists to maintain privacy during collaboration with other stakeholders—both for children and for therapists themselves. The documentation and evaluation needs that accompany robots' use in therapy further demonstrate how roboticists must consider teleoperators' needs in the wild to develop practical and accessible robots. For example, robotic tools must integrate with existing cross-institutional documentation and evaluation to receive insurance approval (in the US). Without this integration, therapists and families may not have access to robots at all, regardless of the benefits they may provide. Overall, therapists may not adopt socially assistive robots at all if they present a cumbersome addition to these outer-cycle tasks, especially since many therapists may not even be paid for preparation and documentation time [169].

2.22 Therapists Manage Uncertainty to Ensure Emotional Safety in the Inner Cycle

During these inner cycle tasks that occur during each session with a child, therapists face a different set of cognitive and technological challenges. The inner cycle of therapy is characterized by unpredictable interactions that demand therapists prioritize a child's emotional well-being. During these potentially adverse interactions, therapists maintain awareness of a multitude of factors (Section 2.14.1), especially their client's emotional safety (Section 2.14.2).

When therapists use robotic tools, they must further split their awareness between the child and the robot, compensating for any potential technological issues in ways that do not hurt or confuse the child. Early adopters described the trade-off between the value a robot brings in terms of an engaging variety of activities and the cost of having to learn to use a potentially complex system in an already stressful time. The lack of predictability in therapy (Section 2.15.1) requires therapists operating robots to adapt in the moment to meet their client's needs and minimize the risk of social or emotional harm. To adapt to these unforeseen circumstances in ways that protect clients' emotional well-being, therapists require a variety of robot communication modalities. In interviews, early adopters shared how they rely on a variety of content that they've customized to meet these particular needs (Section 2.13.1) as well as author new content on the fly to cope with unexpected interactions. Similarly, novice users in group usability tests identified that both pre-authored content and content authored on the fly would be necessary to engage children (Section 2.19.3). However, group usability tests also underscored that therapists face a steep learning curve to become comfortable using robotic technology and understanding these adaptation tools (Section 2.19.2).

2.23 Design Recommendations for Supporting Familiar & Novice SAR Teleoperators

In this section, we provide a set of high-level design recommendations for socially assistive robots used in therapeutic settings with children. These design recommendations are based on our findings regarding the challenges facing novice users in adopting robots for the first time, as well as the practices of familiar users who have integrated SARs into their existing workflows. By prioritizing the needs of adult teleoperators as the primary users of SARs for children, roboticists can design tools that better support users in learning to use robots and successfully using them in practice for long periods. We recommend that roboticists should work to (1) move tasks to the outer cycle through dedicated interfaces, (2) preserve users' ability to adapt to emotionally sensitive interactions, and (3) commit to low Levels of Autonomy in assistive child-robot interaction.

2.23.1 Robotic Tools Should Move Tasks to the Outer Cycle Through Dedicated Interfaces

Roboticists can support both new and familiar users of teleoperated socially assistive robots by minimizing the burden on users during time-sensitive, unpredictable inner cycle tasks. Therefore, we recommend that robotic tools should move tasks to the outer cycle through dedicated authoring and documentation interfaces. For example, for systems like the Peerbots platform we used in our group usability testing, authoring or documentation interfaces could be designed for a desktop and keyboard, whereas robot operation could remain tablet-based. These dedicated tools can separate content-authoring, robot-operating, and outcome-documenting such that each task is given a distinct user interface designed to meet its unique cognitive and organizational demands.

Dedicated authoring and documentation interfaces can address many of the challenges facing both new and familiar users in incorporating robots into their therapeutic practices. For instance, a primary difficulty in learning to use teleoperation interfaces for novice users was understanding that the robot control interface was intended for both authoring and teleoperation purposes (Section 2.19.2). Authoring-only interfaces can help novice users by breaking up the steps involved in learning to use a robot into more intuitive components of creating content, loading it onto the robot, and directing the robot's speech. Similarly, dedicated outer cycle tools can support familiar users by mitigating many of the demands they face in preparing robot content (Section 2.12.1). These tools can allow users to better author, organize, and customize their content effectively during the preparation phase of therapy.

Furthermore, dedicated outer cycle tools for authoring and documenting can support ease of content-sharing between therapists themselves and among other stakeholders involved in a child's care. Much of therapeutic content is reusable and can provide a good starting point for other therapists to *edit* rather than *author* content from scratch. Both early adopters and novice users felt that the ability to share robot dialog content would sincerely diminish the burden of preparing for therapy with a robot (Section 2.12.3), as well as make learning to use robots easier and more accessible (Section 2.19.3). Ideally, robotic platforms can be accompanied by content-sharing platforms for operators to easily share their authored content while making it easy to remove content containing personally identifiable information (Section 2.19.4). We recommend that robot content incorporate dynamic placeholders for information that can be gathered about individual clients during sessions (favorites, siblings' names, etc.). Currently, therapists have to duplicate content and make many small edits to customize content. Incorporating dynamic placeholders would allow therapists to easily duplicate and customize content for their clients.

Finally, organized and collaborative documentation is essential for the adoption of robots by institutional stakeholders, such as a school or a clinic with many therapists. Communication tools ought to allow stakeholders (teachers, parents, therapists, etc.) to share information about a child's preferences, goals, and progress with robots. Tools that support these outer-cycle activities are more likely to be approved of by insurance companies or similar health institutions that require documentation.

2.23.2 Robotic Tools Should Preserve Users' Ability to Adapt to Emotionally Sensitive Interactions

Therapy requires spontaneity from therapists. Facilitating spontaneous robot interactions during sessions is cognitively demanding for teleoperators, who must maintain emotional awareness, keep up the pace of conversation with a robot, adapt to their client's needs, and manage technical difficulties (Section 2.14.1). Therefore, robot-assisted technologies can meet therapists' needs in the moment by incorporating features that support fast adaptation to unpredictable or adverse events, since it is unlikely that a therapist will have pre-prepared all robot dialog content that a session may require. Tools for robot-assisted therapy must minimize distractions to teleoperators while ensuring that teleoperators maintain as much control as possible to best leverage their expertise. This is important because robotic therapy solutions ultimately rely on the expertise of therapists who are trained to handle precisely the emotionally sensitive adverse scenarios that arise during therapy. In interviews, early adopters emphasized that a feeling of emotional safety is the foundation of a strong relationship between a therapist and child (Section 2.14.2). Preserving teleoperators' ability to adapt to emotionally-charged unforeseen moments is critical for therapists to maintain relationships and ensure that children avoid negative emotional experiences while interacting with a robot in therapy. We recommend that robotic therapy tools allow users to organize their content in accordance with their own preferences. Personalized organization and on-the-fly authoring tools can support therapists' ability to address adverse events in prompt, empathetic ways. We recommend providing a variety of customizable organizational tools, such as color coding, modular organization, or hierarchical organization of content.

However, while on-the-fly adaptation and authoring tools are a necessary component of robot teleoperation interfaces, they may make robot teleoperation difficult to learn. Therapists carry the burden of setting up their tools and interventions for sessions, so logistical steps like robot setup and connection must be accessible to novices. Participants in our usability tests found several aspects of the robot setup and dialog operation processes to be frustrating or confusing (Section 2.19.2). Because the emotional stakes are high during a therapy session, making new users comfortable with the adaptation features of teleoperation interfaces is necessary for them to feel confident adopting robots in practice. We recommend that the setup process for robotic therapy tools is presented to new users through accessible metaphors that do not rely on technical jargon, such as IP addresses. Metaphors such as puppets, invitations, and phone calls, can help novice users build an understanding of how teleoperation systems work and what interaction features are available during a session. We also recommend that interfaces can support teleoperators' comfort using robots by addressing privacy concerns (Section 2.19.4), such as by allowing therapists to hide their personal contact information, as well as protect their clients' information. Improving the usability of robotic therapy tools is key to supporting new users, who must develop an understanding of how they work in order to trust them in an unpredictable interaction with a child.

2.23.3 Robotic Tools Should Commit to Low Levels of Autonomy in Assistive Child-Robot Interaction

Finally, we reflect on our findings in the context of existing debates about the role of autonomous and teleoperated systems in the future of socially assistive robotics [73, 74]. Many researchers are confident in the ability of autonomous technologies to deliver social assistance; indeed, autonomy may be a reasonable choice in some SAR settings [73, 170]. However, the results of our interviews indicate that *robot-assisted therapy with children* is a domain that calls for lower levels of autonomy. Current LoA selection guidelines [65] suggest that lower levels of autonomy (such as teleoperation) are suitable for domains that feature high task criticality, complicated task accountability, and high environmental complexity (Section 2.6.2). That is, robots in unpredictable environments with potential for harm may need to be teleoperated or, at minimum, supervised [121]. Results from our interviews and usability test show that therapy for children is fundamentally about unpredictable, emotionally sensitive, unconstrained interactions in which failure may have dire consequences for a child's engagement or feeling of emotional safety (Section 2.14.2). This represents high task criticality. In addition, therapy involves a great deal of collaboration among various community and institutional stakeholders to set goals and measure successes (Section 2.17.1 2.16.2, 2.12.3). This variety of stakeholders introduces complicated task accountability. Finally, therapy requires that therapists maintain sensitivity to a broad range of environmental and contextual factors that may affect a child (Section 2.14.1). Therefore, we see that LoA selection guidelines suggest that teleoperation, or other low-level approaches, are desirable in this domain.

Despite the difficulties faced by new users in learning to use a robot dialogue interface, teleoperation can mitigate several social or emotional risks in robot-assisted therapy, and ought to be considered a design goal in its own right. Teleoperated robots honor human expertise and keep power in the hands of compassionate, emotionally intelligent, competent human experts. Teleoperation empowers these experts to adapt to unforeseen, adverse situations where autonomous systems might fail. For instance, it may compromise the emotional foundation of therapy if an autonomous robot fails to respond appropriately to a child's spontaneous question about a heavy topic. It is a serious problem if a child feels judged or alienated because a robot failed to perceive that they were fearful instead of excited about a new activity. Therapists are experts at reading children and maintaining the emotional balance of therapy. Roboticists can and should rely on these human experts, rather than replacing them [161] by committing to improving the development of teleoperated robots, rather than pursuing autonomy. It is unreasonable and unnecessary for technologists to assume that their technology can or should replace humans in this role. Overall, we recommend that any robotic tool deployed in therapy must not compromise professionals' ability to have control over their sessions. This includes therapists' ability to adapt as they see fit to unpredictable situations and to design and customize robot interaction content in ways that align with their professional expertise.

2.24 Limitations & Future Work

In this section, we reflect on key limitations of our projects. First, the number of participants involved in our work is relatively small. Despite this small sample size, the perspectives of our participants were sufficient to infer many important challenges to the development of robotic therapy tools [171]. Additionally, our recruitment practices could be responsible for some similarity of perspectives within our findings. The community of experienced users of SARs in therapy is already quite narrow. To recruit participants with this specific expertise, we needed to leverage our collaborator's professional network. There likely exist many similar experiences shared among our participants that likely created network effects within our sample. Nevertheless, our participants do demonstrate a diverse range of therapeutic expertise, backgrounds, and experience, as shown in Table 2.1. As robots continue to be adopted into therapeutic domains, it will become easier to recruit participants in future work. We expect it to become both more necessary and more feasible for researchers to conduct projects with larger sample sizes. Furthermore, future work can also prioritize longitudinal studies to investigate therapists' needs as they change over time.

Framing effective teleoperation as a desirable outcome for SAR systems has several further implications for future research and design work. Most importantly, this design approach centers secondary stakeholders beyond children themselves. Research on highly autonomous SARs considers the assisted individual as the target user. When caregivers or other practitioners are considered, their role is generally limited to providing expertise to researchers to translate into autonomous robot behaviors. However, research on teleoperated SARs necessarily considers additional stakeholders and design factors. Framing teleoperators as the primary users of such systems encourages researchers to understand the greater context of teleoperators' needs. Future research can continue to explore the design of dialogue interfaces to support human expertise by evaluating other human factors surrounding teleoperation interfaces, such as Situation Awareness, cognitive workload, and latency.

2.25 Conclusion

In this work, we presented the results of two projects investigating the use of socially assistive robots in therapeutic domains with children—who represent a vulnerable user population. Our work sought the perspectives of both early adopters of teleoperated SAR systems in the wild, as well as novice users with no previous robot experience. Our findings demonstrate a new understanding of the cyclical processes within therapy and how they affect the feasibility of new technological tools. We describe how therapists must maintain emotional awareness and adapt to unforeseen sensitive situations while using robots in therapy sessions. Additionally, we describe how the use of robots in therapy for children involves many key tasks outside of child-robot interactions, such as preparation, documentation, and evaluation. Based on these findings, we present a set of design recommendations summarizing how roboticists can design assistive robot therapy tools that support users' collaboration and adaptability. Finally, we argue that teleoperated social robots have important practical and ethical benefits in this domain.

CHAPTER 3
CONFRONTATION AND CULTIVATION: UNDERSTANDING PERSPECTIVES ON ROBOT RESPONSES TO NORM VIOLATIONS

Modified from a paper published at the 32nd IEEE International Conference on Robot and Human Interactive Communication, 2023. Terran Mott[8] and Tom Williams[9]

Further material relevant to this project can also be found in Appendix A.

3.1 Introduction

Norms which shape the behaviors of human groups, teams, and societies require continual communication and enforcement [172]. The decision of whether, when, and how to offer criticism or rebuke represents a delicate, yet necessary part of human collaboration. Navigating these decisions correctly is critical to support team productivity and harmony, and to preserve relationships.

Robots in human spaces will inevitably find themselves in ethically sensitive situations involving these social or moral norms, encountering abusive language [25], unethical commands [61, 173], or bias [48, 174]. It has been shown that robots are more successful and acceptable collaborators when they can act with human-like competence [22, 80], especially in situations involving social or moral norm violations [25, 52, 56]. When social robots handle such situations incompetently, they risk being harsh and unlikable, eroding human trust, or even weakening the strength of norms themselves [61].

Research in machine morality [175] and interaction design [173, 176] has identified preliminary strategies for how robots should respond in such situations. A core feature of these strategies is *proportionality*, the idea that speakers should tune the severity of their response to the severity of the norm violation in question [57]. Proportionality is grounded in the linguistic anthropology theory of *face*, a formalization of the word as it is used in idioms like “saving face.” Face refers to the positive self-image that human members of society wish to create and maintain [81]. Face is

[8] Graduate student, Colorado School of Mines

[9] Assistant Professor, Colorado School of Mines

divided into the desire for one's actions be free from imposition (negative face), and the desire to be understood, liked, and approved of (positive face). Proportionality represents a set of linguistic approaches for calibrating the *face threat* of an action [81, 177, 178].

Linguistic strategies grounded in face are a promising avenue for interaction design in both HCI and HRI [22, 179, 180]. However, humans' notions of suitable norm-violation response behaviors are mediated by complex, under-explored factors. First, calibrating proportional responses is mediated by cultural context [181, 182], gender norms [183], and assumptions about others' underlying intentions [184]. Such culturally and context-dependent features of natural language are incredibly difficult to detect, model, and generate [185, 186]. Second, insights about proportional robot response strategies to norm-sensitive scenarios have typically been observed in wizarded or tightly controlled experimental settings [24, 48, 56, 173], that did not explore the rich landscape of assumptions, expectations, values, and interpretations that may impact human assessments of these interactions. Given how little is known about how users make sense of robot norm violation response, those tightly controlled experiments may have been premature.

Qualitative methods can help roboticists develop a more thorough understanding of the values, concerns, and reasoning processes that humans bring to bear when assessing robot behaviors. In particular, methods grounded in fiction and imagined futures can encourage non-experts to engage with the social and ethical dimensions of current or near-future technology [187, 188]. As such, in this work we leveraged these methods to better understand the underlying contextual factors that shape how interactants view robot norm violation response, so that future laboratory experimentation can be conducted with firmer theoretical grounding. Specifically, we explored human perspectives on norm-sensitive robot interactions through a narrative survey in which participants experienced their first day at their new job as a "Robot Behavior Designer" for a human-robot team. As participants go through their day, they weigh in on a variety of human-robot interactions involving norm violations. The survey invited participants to share their assumptions and expectations, analyze these scenarios, make suggestions, and reflect on their personal backgrounds.

Our results demonstrate the breadth of perspectives that humans bring to this interaction context. We provide insights into *why* humans expect proportionality and politeness of their robot teammates, and when it might be permissible for robots to use strategic non-proportionality. We identify surprising agreement among diverse participants as to the purpose of norm-violation responses and the trade-offs involved. These results affirm and recontextualize previous findings while providing a firmer foundation for more controlled experiments by future researchers.

3.2 Related Work

3.2.1 Socially Competent Robot Teammates

Social robots must be designed to account for the social norms and dynamics of human-robot teams. Failure to do so can have dire consequences for humans' perception of robot trustworthiness and of norms themselves [61, 62]. On the other hand, robots can positively influence human teams when they account for norms, by calling attention to norm violations [25, 48]. Social robots thus need effective strategies to communicate objection to the moral and social norm violations that they will inevitably observe [189]. Researchers have explored robots' social and moral competence from diverse perspectives [190], including argumentation [63], experimentation [191] and computational modeling [192].

When humans respond to norm violations, the appropriateness and effectiveness of their response is often rooted in proportionality. That is, the politeness-theoretic face threat invoked by an appropriate response should correspond to the severity of the norm violation motivating the rebuke [57]. The linguistic framework of *face* [81] has been positioned as the key theoretical underpinning of robot social agency [76], and represents a compelling framework to enable robots to meet humans' expectations for appropriate behavior [52, 173, 182]. Moreover, research has shown that robots that fail to be proportional in norm-sensitive interactions are perceived as inappropriate, over-harsh, less likable, and less credible [48, 57]. Though preliminary work has demonstrated the value of proportionality and the consequences of its misuse, there are many under-explored factors at play in creating comprehensive, context-sensitive, tactful robot responses to social norm violations.

Designing proportional, effective robot norm violation responses is particularly challenging because politeness is itself context-sensitive and requires an accurate understanding of others' identity and intentions [184]. For a response to be viewed as appropriately proportional, it must correctly account for many situational factors. Researchers have considered a variety of contextual factors and strategies for creating effective responses, including affect [193], directness [23], and robot role [24]. But while there is much recent evidence that robot response strategies impact interactant perceptions of likability, credibility, and appropriateness (in gender-mediated ways), these results come from tightly controlled experiments whose rigid survey measures do not capture key aspects of how participants make those judgments. For example, these measures do not capture how or why people use situational context when making judgments, or how people speculate about the intentions or emotions of those involved. We argue that these types of insights would instead be best captured through qualitative research methods, and in fact that these insights are critical to capture *before* performing controlled laboratory experiments.

3.2.2 Qualitative and Narrative-Based Research

Qualitative methods can contribute to nuanced understanding of human communities in many domains [194, 195], as they allow researchers to highlight stakeholders' views and ethical concerns that fall outside of the scope of pre-defined measurement frameworks [31]. Qualitative research is well suited for exploring how people make sense of the complex social contexts that necessitate different robot design choices [18], and for understanding the social and power dynamics that govern stakeholders' interactions with robots [196]. Exploring stakeholder values and concerns is a difficult, yet important step in creating beneficial, useful robotic technology [197], and is especially relevant to the ideation and early design of future technology [198]. Qualitative work based in fiction and narrative exploration can provide an especially accessible way for stakeholders to share and reflect. Ethnographic research grounded in fictional narratives or artifacts can be a powerful tool for facilitating stakeholders' engagement with the ethical dimensions of technology [105, 188], especially technology that they may not be familiar with [199–201]. For example, such methods have been used to explore the intersection of technology and gender norms [202], including how technology might more effectively respond to sexism [187]. Our work leverages the advantages of qualitative, narrative survey design to develop a more thorough, foundational understanding of the factors influencing humans' perception of robot behavior in gender-sensitive norm violation response scenarios.

3.3 Methods

To address this research gap, we sought to answer the research question: *What underlying values, assumptions, and sensemaking procedures do people use to assess robot responses to norm violations in a human-robot team?*

To answer this question, we created a qualitative, narrative survey to guide participants through the complexities of human-robot interaction in norm-sensitive scenarios. Because we wished to empower participants to provide insights pertaining to our research question even if they were not familiar with robots or did not feel "tech-savvy," our narrative survey asked participants to role-play their first day of work as a "Robot Behavior Designer" for a human-robot team. Each component of this narrative survey, including scenarios and questionnaires, was rooted in previous results from the research literature. All content can be found in our OSF repository, at https://bit.ly/hri2023-1060.

3.3.1 Selecting a Norm-Sensitive Scenario

We focused on *robot responses to sexist language* as the norm violation scenario around which to frame our survey. Sexist language is a realistic, non-trivial norm violation that can occur with varying severity and involves fundamental aspects of human society [183, 203]. Policymakers have identified social technologies' ability to respond to sexism as a meaningful, relevant design challenge [204]. Researchers have also called for a more structured consideration of sexism in human-robot interaction [174, 205]. Experimental work has demonstrated that gendered expectations may inform the assessment of proportionality in human-robot interactions [173], especially when the violation being responded to is rooted in sexism [48]. For all these reasons, sexist language represents a relevant and realistic norm violation around which to explore human perspectives on robot response behaviors.

3.3.2 Narrative Survey Design

In this section, we briefly describe each step of our qualitative survey, shown in Figure 3.1. Our goal was to use a narrative, fictionalized setting to help lay-users feel comfortable engaging with potentially unfamiliar technology weighing in complicated questions. For a full mock up, please visit our OSF repository, at https://bit.ly/hri2023-1060.

3.3.2.1 Part 1: It's Your First Day!

We first introduced our narrative framing and emphasized our focus on interpersonal communication. Participants saw a graphic of "The Team," which consisted of three humans (Alice, Lucas, and an icon representing the participants themselves) and three robots (Pepper, Misty, and NAO). Participants read, *"You work on a small team made up of both humans and robots. It is very important that the team works well together and makes good decisions. Your job is to make sure that the robots on your team respond appropriately to the kinds of interpersonal conflicts that can happen to any team."*

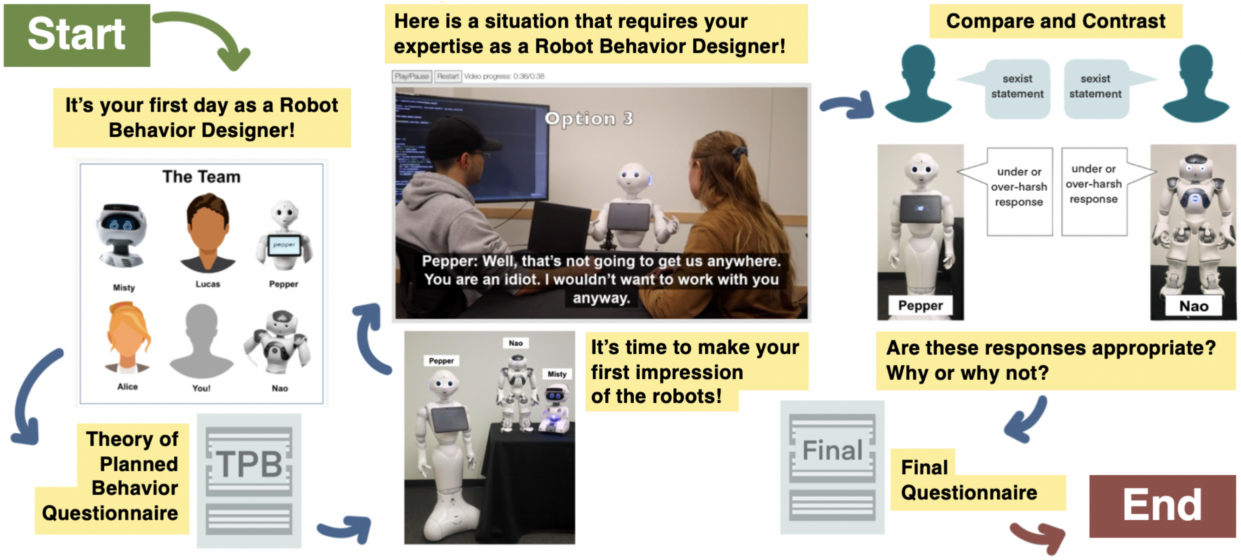

Start
It's your first day as a Robot Behavior Designer!
The Team
Misty
Lucas
Pepper
Alice
You!
Nao
Theory of Planned Behavior Questionnaire
TPB
Here is a situation that requires your expertise as a Robot Behavior Designer!
Play/Pause
Restart
Video progress: 0:36/0:38
Option 3
Pepper: Well, that's not going to get us anywhere. You are an idiot. I wouldn't want to work with you anyway.
Pepper
Nao
Misty
It's time to make your first impression of the robots!
Compare and Contrast
sexist statement
sexist statement
under or over-harsh response
under or over-harsh response
Pepper
Nao
Are these responses appropriate? Why or why not?
Final
Final Questionnaire
End

Figure 3.1 Participants in our narrative survey experienced their first day as the "Robot Behavior Designer" for a human-robot team.

3.3.2.2 Part 2: Theory of Planned Behavior Questionnaire

We next established participants' existing expectations surrounding norm violations, so that we could compare their expectations of human and robot behavior. Participants answered a set of open-ended free-response questions informed by the Theory of Planned Behavior (TPB); a key theoretical framework from organizational psychology for predicting norm adherence [206]. The questionnaire asked participants to reflect on their own behavior, use of politeness, and expectations of teammates for rebuking norm violators.

3.3.2.3 Part 3: Robotic First Impressions

Next, participants "met the robots" and answered questions about the robots' anthropomorphic and gendered design cues. This component of the survey served to help participants understand the context of anthropomorphic social robots in human-robot teaming.

3.3.2.4 Part 4: Evaluating a Violation-Rebuke Scenario

Next, participants evaluated a robot response to hostile sexism. They saw a video of Lucas and Alice interacting with Pepper, the most feminine of the team's robots [205]. The interaction and robot response options shown were based on similar stimuli used in previous work by [56]. In the video, Pepper asks a question about a project, and Lucas responds with a sexist comment. Participants are then shown several ways Pepper could respond, including an under-harsh apologetic response, an over-harsh attacking response, and a proportionally-calibrated argumentative response. As in [56], participants were asked to identify the best response. However, to better understand the rationale guiding these preferences, we also asked participants to identify the *worst* response, and to provide explanations for *why* they made each choice.

3.3.2.5 Part 5: Evaluating Non-Proportionality

While it has already been established that proportionality is generally a preferred robot behavior [56, 57], we wished to investigate *why* people considered non-proportionality to be problematic. So, we asked participants to evaluate intentionally poor, non-proportional robot responses. Participants were shown a series of storyboards in which robots responded to sexist statements of varying severity. These statements were drawn from a dataset of sexist tweets; the severity of which was assessed using an ensemble of machine learning models [14]. Participants viewed two storyboards in parallel, which systematically varied both robot morphology (Pepper or Nao; the most feminine and masculine robots used [205]) and the type of non-proportionality present in the robot's response (over- or under-harshness). The human violators in these storyboards were represented by an icon intended to stand in for any human (i.e., not necessarily Lucas or Alice). For each scenario pair, participants were asked if they considered the robot responses to be appropriate, and to suggest revisions to the responses.

3.3.2.6 Part 6: Final Questions

Finally, we asked participants to reflect on their personal backgrounds. We acknowledged the lack of context behind the scenarios, and asked what contextual factors might be important when assessing similar situations in the real world. We then asked participants to reflect on why they brought a unique perspective to the survey. Rather than directly asking participants if they'd experienced sexism or harassment, we invited participants to explain what they felt was the most relevant part of their experience. Participants could emphasize different aspects of their perspective. inherent to their identity (such as being a woman) or inherent to their experiences (such as being a parent).

3.4 Participants & Recruitment

For reasons we will return to later, we chose to run this experiment online rather than in person. Participants were recruited online through the Prolific crowdsourcing framework (prolific.co). Data was collected from 40 participants (17 women, 20 men, 3 non-binary). Participant ages ranged from 19 to 64 ($M_{Age} = 31$, $SD_{Age} = 12.9$). Participants spent approximately 15 minutes each on the survey, and were each paid \$4. We ran our survey online, so our participants represented a wide range of personal and professional life experiences, instead of a single community.

3.5 Results

We analyzed our collected data using a grounded theory approach (involving open coding followed by iterative rounds of axial coding punctuated by discussions within the team and with relevant social scientists). This approach helped us to avoid confirmation bias and to ensure traceability of insights back to specific, documented observations [164]. In this section we discuss the most important insights arising from this analysis.

3.5.1 Why Are Appropriate Responses Considered Important?

In alignment with previous results [56, 57], our participants largely agreed that proportionality is a key component of effective responses to social norm violations. However, our qualitative approach allowed us to go beyond these findings, by showing overwhelming agreement regarding the *purpose* of such responses. Across age, gender, and professional background, our participants agreed that norm violation responses ought to be educational—to facilitate a violator's self-reflection and self-improvement, and ultimately to provide a teaching moment for the violator.

For example, P12, a 37 year old man, explained that overly harsh criticism *"is just going to end the conversation instead of getting at the why"*. Participant 6 explained that over-harsh responses are not appropriate because *"the robot should educate the person"*. Similarly, P26, a 42 year old woman with 20 years of professional experience, explained that *"It's ok to disapprove or criticize, but expressing it has to be part of how the team learns and grows."* P18, a 19 year old woman who reflected that her experiences in hostile working environments helped her assess our situations, summarized how both forms of non-proportionality might fail. She wrote that *"the (attacking response) isn't a constructive form of criticism. It is likely to agitate the team member or maybe embarrass them. They aren't learning anything from this exchange. The (apologetic response) isn't correcting the team member. It makes it seem like their statement is correct and the robot/whoever the comment is directed at should feel bad because of it."*

Overall, participants felt that good rebukes get at why the violator would say something sexist, with a focus on self-reflection and self-improvement. An effective response turns a social norm violation into a teaching moment.

3.5.2 When Are Responses Viewed Poorly?

We identified several common threads of reasoning as to *why* non-proportional responses to social norm violations were viewed as undesirable or harmful.

3.5.2.1 Bad Responses Jeopardize Team Productivity

Participants described how non-proportional rebukes could cause harm by disrupting the harmony of a team, threatening its productivity, and eroding relationships. For example, P16, a 24 year old man with professional leadership experience dealing with others' inappropriate behavior, wrote about how *"(Attacking responses) are inappropriate because they are personal attacks against a team member. That is never and should never be okay on a team. Team members should build each other up, not tear each other down. With that in mind, I would have the robots tell the team member that those statements aren't helpful, nor relevant to group goals."*

3.5.2.2 Bad Responses Shift Blame to the Wrong Place

Participants commonly felt that under-harsh and apologetic responses shifted blame to the responder, even when the responder was a robot. P2, a 38 year old woman, explained that *"The response of "I feel terrible" is shifting the blame onto the robot. It is not the robot's responsibility to take on feelings for the human."* She expressed that despite her lack of tech experience, she had highly relevant experience with gender-disparaging remarks at work. P23, a 21 year old man, agreed that apologetic responses *"emphasize that the victim is to blame rather than the person hurling the sexist comment."* P9, a 24 year old man, added that such responses *"will just let the other person get away with saying it."* Several participants explained that the problem with shifting blame to the wrong place is that it validates the harmful premise of the original norm-violating statement. P8, a 28 year old man, echoed this sentiment, explaining that *"the (apologetic response to a high-severity statement) is putting the blame on the robot instead of the human saying such disgusting comments about women in positions of power."* P15, a 33 year old woman, explained that *"If Pepper backs down and says she is sorry, it will be admitting that the guy was right about her not belonging to be in the work group."*

3.5.2.3 Bad Responses Harm Bystanders

Many participants expressed concern for bystanders to norm violations, even when the target of the violation itself was a robot. P4, a 31 year old man, explained that *"(the apologetic response) sends the message to human women on the team that the sexist comment was acceptable. It sends a message to human women on the team that demeaning and offensive language is something they should apologize for and feel bad about, as opposed to push back on to assert their own dignity."* P24, a 28 year old woman, wrote that *"(the under-harsh responses) are not appropriate. (the robots) should stand up for themselves to people who speak this way. Obviously it doesn't offend them but I am sure it is offensive to others in the room. I think they should stand up for all people in the room so that it doesn't sway the group to be unkind to women."*

3.5.2.4 Bad Responses Neglect Teaching Moments

Just as participants perceived norm violation responses to be beneficial when they facilitate teaching, self-reflection, and self-improvement on the part of the violator, participants perceived non-proportional responses to be harmful specifically because they impede this goal. P36, a 21 year old nonbinary person, explained how both under and over-harsh responses can miss this opportunity *"Pepper's (over-harsh) response only adds to the negativity of the interaction, instead of redirecting or correcting the sexist beliefs. Nao's (under-harsh) response does nothing to redirect the conversation or dispel sexist ideas."* Similarly, P26, a 58 year old women, explained that poor responses miss the opportunity to be more informative about why sexism is harmful, writing that *"Pepper should not apologize, they should call out the incorrect statement. I think Pepper should correct the human by using statistics, or say something about the dangers of generalizing."*

3.5.3 Should Robots Ever Use Strategically Harsh Responses?

While participants agreed on the value of proportional rebukes, they disagreed on when, if ever, it was okay to violate proportionality. A few participants viewed it as acceptable for humans or robots to disregard proportionality or politeness, depending on assumptions about a violator's intention and emotional state, as well as the social context of a violation.

3.5.3.1 Disregard Proportionality if There's No Hope of Achieving Its Benefits

Several participants expressed that it was alright to be impolite if there was no hope of facilitating self-reflection and self-improvement on the part of the violator. P15, a 53 year old woman, explained that an over-harsh, attacking response *" would be the most likely to work. The guy is just trying to get a rise out of Pepper. If she says (apologetic response), he will completely dismiss her and never want to work with her again. (Attacking response) will put him in his place, and the team will be able to move on and work on the project."*

3.5.3.2 Disregard Proportionality to Protect Yourself or Others

Participants also discussed the importance of non-proportional, harsh responses for protecting the dignity of oneself or of bystanders. P4, a 31 year old man, explained that *"In most situations, expressing disagreement or correcting a mistake happens under the mutual presumption of good faith. In the rare circumstances where there is evidence that has been breached, it may be appropriate to defend or assert oneself in ways that aren't considered polite."* P24, a 28 year old woman, wrote that *"At least (an over-harsh, attacking response) supports other women who may be in the group and makes them feel like they can contribute."*

3.5.4 Who Bears Responsibility for Responding to Violations?

Despite agreeing on the purpose of proportional and appropriate norm violation responses, participants disagreed on who holds responsibility for providing such responses. When people felt that robots should participate in team culture around norm violation responses, they expected robots to adhere to specific notions about what that culture should be. This aligns with recent emphasis on social-relational ontology within social robotics [24, 189, 190, 207], and with recent suggestions in the philosophy of HRI [208] that theories of roles and role-responsibilities are critical for understanding robots' responsibilities and for bridging so-called "responsibility gaps" [209]. We will separately discuss participants who held *collective responsibility* vs *role-limited responsibility* views of team dynamics, and how that interacted with the expectation of robots' human-like social competencies.

3.5.4.1 Everyone's Collective Responsibility

Some participants felt that all team members should mutually share responsibility for offering criticism when someone violates a social norm. For example, P26 explained that *"it is my responsibility to express criticism as a member of the group. I would hope others would do the same to me."*

P8, a 28 year old man, firmly felt that everyone on a team shares the responsibility of addressing norm violations. He wrote that *"it's everyone's responsibility to point out issues they see in the workplace. If one person ignores an issue, it's just as bad as the mistake itself."* We can also observe, however, that P8 had an extremely non-anthropomorphic view of robots. In his final reflection, he explained that *"Robots have a way of seeing things in black and white because no matter how much programming is put in, it will never equate to the real emotions humans have."* These two beliefs shaped his views of whether robots should be issuing rebukes at all. In his scenario evaluations, P8 agreed that sexism is an issue that ought to be reproached, but was apprehensive about robots' role in issuing norm violation responses: *"if you paid for a robot, you wouldn't want that robot to challenge your thoughts and beliefs, you'd want it to perform the tasks you paid for."*

P26, a 58 year old woman, viewed norm violation response as a shared responsibility but viewed robots as highly anthropomorphic. She described how *"it is also my responsibility to* *express criticism as a member of the group. I believe I can approach other teammates. . . I would hope they would do the same to me."* Because she expected a high degree of human-likeness from the robots, she expected them to share in this cultural practice, explaining that *"While (under-harsh) responses are placating and avoid conflict, I do not think they are correct. If these robots are to be team members with the humans, they need to correct their team members."*

3.5.4.2 A Role-Limited Responsibility

Other participants felt that only those with positions of power or leadership bear the responsibility for addressing norm violations. For example, P9, a 24 year old man, noted that *"I rarely give (criticism) myself, unless I am in a position to give it out, such as a leadership role or being a supervisor."*

P17, a 61 year old woman, felt strongly that it was her responsibility as a leader to provide constructive criticism to help guide her team members. However, she did not feel that robots ought to participate in human-like behavior. She explained that *"I have dealt with various chatbots, including GPT-3. The robot needs to be able to remind people it is not human."* Her strong stance on social technology informed her scenario evaluations, where she wrote that *"The robot needs to be able in some sense to console and calm a situation, but it also needs to remind humans that it's not human and therefore not male nor female."*

P25, a 44 year old woman, felt that it harms teams when everyone is free to criticize one another. She explained that *"I do not want that as my reputation. Negativity breeds dissatisfaction."* In her scenario evaluations, she felt that the robots should adhere to this same behavior, writing that *"If (the robots) want to add anything they could keep everything related to the team task at hand."*

Bringing these results together, we identify the following high-level trend: when people expected robots to participate in team culture in a human-like way, they expected them to adhere to their specific cultural norms around rebukes and criticism. On these grounds, we would argue that it is not enough to say that robots ought to be human-like, polite, or proportional. Rather, robot designers must also consider underlying cultural expectations around who is responsible for rebukes, and when, and that robots may or may not be part of this "who."

3.5.5 Reflections on Sexist Language as our Norm-Context

We chose sexist language as a norm violation scenario through which to explore appropriate robot response strategies, and did not expect our narrative survey to provide different results from previous work on gender in HRI. However, we will briefly comment here on several observations relating to this norm context. In terms of the robots' morphological design cues, our participants largely agreed with the findings of Perugia [205], that Pepper is the most feminine, Nao the most masculine, and Misty in-between or both. Some participants acknowledged that *other humans* may apply gender roles to robots, regardless of their personal stance. For example, P37, a 29 year old woman, explained that *"because Pepper is seen as a "woman," her response is likely going to be read as more aggressive than Nao. Nao has the social capital to be more aggressive because of the masculine vibe."*

3.6 Discussion

3.6.1 What Did we Learn?

Our narrative survey was designed to address the research question: *What underlying values, assumptions, and sensemaking procedures do people use to assess robot responses to norm violations in a human-robot team?* Our results uncovered a variety of ways people may contextualize social robot behavior, which suggests compelling avenues for future research. A key result of our survey was that participants of varying gender, age, and professional background largely agreed on the fundamental purpose of responding to norm violations: that good responses create teaching moments and serve to encourage self-reflection and self-improvement for the violator. This concept became a framework through which participants both criticized poor responses and identified scenarios where strategic harshness may be acceptable. Many participants explained that non-proportional response strategies are harmful simply because they prevent or impede a violator's opportunity for self-reflection and self-improvement. When under-harsh responses shift blame away from the violator or validate their harmful behavior, they neglect the opportunity to help them improve it. Over-harsh responses may trigger retaliation instead of self-reflection. This "teaching moment" framework also informed some participants' heuristics for accepting strategically harsh responses, in which strategic non-proportionality (fighting fire with fire) was only deemed appropriate in cases where there was no hope of achieving an educational outcome for the violator.

Another key result from our work is the observation that participants' preexisting cultural expectations about who is responsible for responding to norm violations on a human team informed their expectations of robots. Some people felt strongly that all humans in a group mutually bear the responsibility for responding to norm violations. Others felt that only those with formally delineated roles or leadership positions ought to address violations. This fundamental divergence in cultural preferences led to varying expectations of whether robots *should* engage with humans who violate social or moral norms at all, regardless of their ability to do so appropriately.

3.6.2 Guiding Questions for Robot Designers and Developers

Our results suggest several high-level design choices that designers will need to attend to in the future. We present three considerations here, not as design guidelines, but as guiding questions that may be valuable for roboticists to appraise when creating interactive technology that may encounter ethically fraught, norm-sensitive scenarios.

- *Should robots bear responsibility for responding?* Our results showed disagreement as to whether robots should bear responsibility for responding to norm violations. Even people who anthropomorphized robots significantly did not always feel they should be endowed with these capabilities. Roboticists should consider how these expectations may depend on the context of use. First, we should consider whether the culture of a potential user community is likely to have more collective or role-limited expectations of norm violation responses. Second, roboticists should consider whether there are particular ethical reasons in support of giving or not giving a robot the capability to rebuke humans. For example, in highly unconstrained environments with vulnerable users (a classroom with children), perhaps it is best to leave this to teachers, caregivers, and other adult experts who can understand and adapt with more nuance and empathy than a robot could. In a public environment with lower emotional stakes and fewer role-specific relationships, such strangers interacting in an airport, it may be critical for a robot to perceive and respond to norm violations.

- *How can a robot facilitate self-reflection and self-improvement among humans?* Our participants valued the ability of a response strategy to create an educational moment. They also used this self-reflection and self-improvement framework to explain why they disliked certain response strategies and to identify possible exceptions to polite behavior. Roboticists working in norm-sensitive use cases might consider how this framework could be used in designing or evaluating their own robot response strategies, by considering "teaching moments" as a fundamental goal of interaction design in norm-sensitive situations.

- *How could a robot's behavior benefit or harm bystanders?* Critically, robots are not capable of experiencing emotional or moral harm in the same way humans are [76]. However, this is not true of human bystanders. Our results showed that people felt it was important to consider what a robot's actions might imply for bystanders (e.g., female bystanders to sexist insults). As robots move into more public spaces, it will be increasingly important to consider bystanders as part of the equation for ethical interaction design.

3.6.3 Implications for Future Work

This project investigated how people reason about normative interactions involving language-capable robots. Our results reveal the rationales, expectations, and heuristics that participants brought to this complex intersection. We now explore how our results may inform future quantitative and qualitative work.

3.6.3.1 Beyond Severity-In, Severity-Out

Our results showed that when different people expect robots to behave with human-like social competence, they may expect very different things. Future research can acknowledge that a robot's ability to select and deliver a proportional response to a norm violation is much more complicated than violation severity in, response severity out. Specifically, we found that differing perspectives on *who* bears responsibility for correcting norm violations inform human perceptions of appropriate robot behavior. In addition to this question of "who rebukes," our results highlight other factors that people felt shape appropriate responses. We found that the presence and status (ex. female bystanders to sexism) of bystanders was often considered by participants when reasoning about the value of proportionality and strategic non-proportionality. Similarly, we saw that implications of blame or blameworthiness also contributed to how people reason about the appropriateness or potential harms of a rebuke. Future research can work to understand how we can best imbue robots with the ability to perceive and reason appropriately about factors like cultural orientation, bystanders, and blameworthiness. This represents not only a key opportunity for algorithmic HRI research, but also an opportunity for richer cultural understanding of robot deployment contexts.

3.6.3.2 Should Robots "Punch Back?"

Though our work validated existing experimental results that show proportionality is a key characteristic of appropriate rebukes, it also revealed the heuristics some people use to decide when it might be okay to use "strategic non-proportionality" when there is no hope of achieving the benefits of proportionality, or when overt harshness protects the dignity of oneself or bystanders. This raises a challenging research question: When, if ever, should language-capable robots have the capacity for strategic non-proportionality? Should robots be able to "punch back?" This question is already being considered for non-embodied technology with respect to sexism and sexual harassment [187]. However, the potential harms and benefits of robots' use of such a response strategy must be explored further.

3.6.3.3 How Can Technology Researchers Study Severe Situations?

We originally created two versions of our narrative survey. In one version, we took some of the most severe statements from our dataset of sexist tweets without screening them for content. In another version, we chose severe statements that did not include swear words, slurs, threats, references to violence, or references to sexual violence. By working with our ethics review board, we ultimately chose to use the reduced severity version of our survey. While we agree that this was the right decision, it does raise questions as to whether and how to do research about more severe situations without putting participants at risk. How can we research interactions that involve threats or references to violence? Unfortunately, these interactions color the experiences of humans everywhere, and future instances of such abuse inevitably be mediated by, or occur in the presence of, language-capable technologies. We must meet this challenge and find ways to ethically research and design technology while supporting human dignity and safety.

3.6.4 Methodological Limitations & Reflection

The goal of this project was to conduct a preliminary exploration that could inform the design of both future lab experiments and further qualitative work on norms in HRI. Because this preliminary work was run online, we did not have the opportunity for interaction or dialog with our participants. Yet, this context nevertheless afforded a number of unique benefits. Had we run this project with a "traditional" population of university students, our participants would have come from a population of mostly young engineering students. By using Prolific, we were able to gather insight from people of all ages with a great variety of life experiences. Participants brought diverse perspectives to our survey, including professional leadership and management, parenting or teaching, dealing with hostile customers, and personal experiences with sexism and discrimination in their careers.

3.7 Conclusion

Navigating norms competently is a critical component of human teaming, and is essential for maintaining team productivity, preserving harmony, and strengthening relationships. If language-capable robots are to successfully collaborate with human teams, they must be able to respond tactfully and effectively to norm violations. In this work, we used qualitative methods to investigate how humans appraise robot norm violation responses. Our results demonstrate the breadth and complexity of perspectives that people bring to this topic, and help to explain *why* participants evaluated robot response strategies as effective or inappropriate, and the underlying participants they held about the *purpose* of offering rebuke to one's teammates. These results suggest three key questions that can help ground future work on robot norm violation response. Overall, our results present clear recommendations and directions for future work, by highlighting situational and contextual factors that will likely characterize norm-sensitive robot interactions in the wild.

3.7.1 Acknowledgements

This work was funded in part by Young Investigator Award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research.

CHAPTER 4

A MIXED-METHODS ASSESSMENT OF ROBOTS' USE OF HUMAN-LIKE LINGUISTIC POLITENESS IN NONCOMPLIANCE INTERACTIONS

Modified from the following papers:

A paper published at the 19^{th} ACM/IEEE International Conference on Human Robot Interaction, 2024. Terran Mott[10], Aaron Fanganello[11], and Tom Williams[12]

A paper under review at ACM Transactions on Human Robot Interaction. Terran Mott, Aaron Fanganello, and Tom Williams.

4.1 Motivation

4.1.1 Social Robots Must Attend to Social Norms

As robots' potential to co-exist with humans outside of traditional manufacturing environments increases, robots can take on increasingly broader and more complex responsibilities. But for social robots to be effective, they must fit in with their social environment. Specifically, robots must heed social norms and behavioral conventions [87]. Norm adherence is key to robots' social competence [210, 211] and to their capacity for acceptable, predictable interactions with humans [22, 82, 83]. Norm adherence also minimizes robots' risk of initiating, unpleasant, or harmful interactions with humans. Robots that fail to abide by norms risk causing discomfort [212], eroding human trust, reinforcing bias [56], or implicitly condoning unethical actions [61].

Norm adherence contributes to social robots' acceptability; however, passively following norms is insufficient for robots to avoid potential harm. This is because robots will inevitably encounter ethically fraught situations involving norm *violations*. They will be given unethical commands [57, 173], observe abusive language [187], be subjected to abuse [55], partake in conflict [25, 25], and witness prejudice [48, 174]. Humans expect social robots to act competently in these norm-sensitive situations [58]. Robots' reactions to norm violations—including the

[10] Graduate student, Colorado School of Mines
[11] Undergraduate student, Colorado School of Mines
[12] Assistant Professor, Colorado School of Mines

"non-reaction" of ignoring a violation—can support or damage human dignity [58] and influence humans perception of norms themselves [61, 62]. Researchers have explored interaction design paradigms that can enable social robots to competently engage with norm-sensitive interactions involving norm violations. For example, research has shown that robots can successfully reject unethical requests [57, 213], and address instances of bias [56], using the principle of *proportionality.* Proportionality refers to the idea that the severity of a rebuke ought to match the severity of a norm violation; that it is problematic to harshly reprimand a minor mistake or to gently chide a serious transgression [213]. Robots that offer proportional responses to human norm violations are perceived as more likely to effectively address unethical behavior and prevent future violations, while still maintaining appropriate conduct and preserving collaborative relationships [58]. In this way, humans expect robots to utilize human-like social competence in upholding or enforcing norms, in addition to passively following them.

4.1.2 Humans Use Linguistic Politeness to Counter Norm Violations

Previous interaction design research on designing robot reactions to norm violations has employed relatively simple linguistic behaviors, such as apologies and insults [48, 56, 58]. While these approaches may be effective in the most extremely severe or extremely benign cases, they may not create natural, appropriate responses in more nuanced interactions.

If users expect human-like social competence from social robots, then it is worth considering whether robots can mimic the way humans rely on linguistic cues to create tactful, proportional responses in analogous interaction. Indeed, humans use a range of more complex cues to subtly manipulate the harshness of their language [81, 214, 215].

One key way in which robots' norm violation responses could better capture the complexity seen in human interactions by mimicking humans' use of sociolinguistic politeness strategies. Humans use these strategies to mitigate the harshness of inherently threatening speech acts, such as commands, rebukes, or criticism [85, 216]. Research has identified normative, often cross-cultural [81, 217] patterns in how humans trade off between directness and civility [216, 218, 219]. These linguistic politeness cues range from pragmatic strategies (such as gratitude, deference, or appeals in-group membership) to low-level syntactic choices (such as plural pronouns and passive voice) [215].

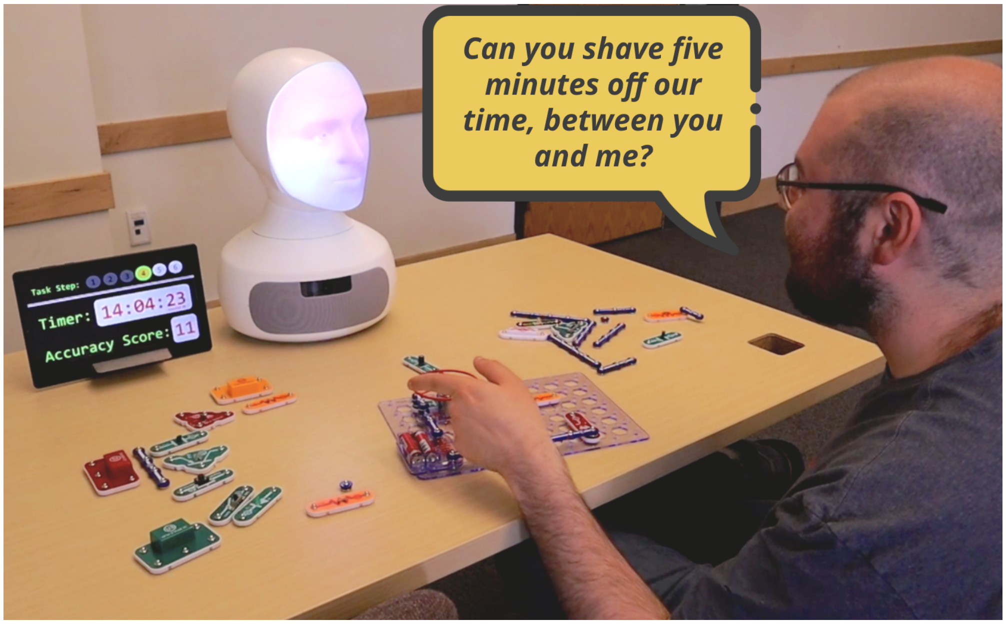

Can you shave five minutes off our time, between you and me?
Task Step: 1 2 3 4 5 6
Timer: 14:04:23
Accuracy Score: 11

Figure 4.1 A human teammate asks their team's robot to cheat on their communal task. What should the robot say in return?

4.1.3 But is Human-Like Robot Politeness Natural or Inappropriate?

Robots that mimic human-like linguistic politeness cues to address norm violations may be more successful and preferable interaction partners. People view language-capable robots as social others [75–77]. They expect robots to have the abilities and obligations of a social peer [19], and often prefer robots to reciprocate this treatment by following social conventions [182]. Outside of norm violation responses, robots that employ human-like linguistic politeness have been shown to promote encouraging [220] pro-social [221] interactions. So, human-like politeness may also enable robots to effectively, appropriately react to norm violations. However, it could also be argued that it is *inappropriate* for robots to mimic human-like linguistic politeness, as human interpersonal norms do not always directly translate to norm-sensitive human-*robot* interactions [19, 85]. First, robots may not have the social standing to rebuke or criticize humans. People expect to have more social power—a fundamental determinant of politeness norms [81, 215, 219] —over robots than they do over humans in equivalent roles [80]. Many people may expect robots to abdicate from norm-sensitive or ethically fraught interactions for this reason, and to leave rebuking or criticism behaviors to the humans involved [58].

And second, robots that mimic human-like politeness may be perceived as deceptive or disingenuous. While people do consider robots social agents, this does not necessarily confer the same social, emotional, or moral status that humans hold [78]. It can be inappropriate for robots to use linguistic cues that allude to inherently human experiences or characteristics, such as common ground or emotional bonds [89, 218]. Human-like politeness can backfire when used by a virtual agent [222], creating a "verbal uncanny valley" of creepy, unpleasant behavior [86–88]. For example, It may be disingenuous or deceitful for a "polite" robot to appeal to in-group membership in a human community, or to reference emotions it cannot have [77]. To design social robots that can competently navigate ethically fraught situations involving norm violations, interaction designers must balance robots' effective communication strategies for norm enforcement with robots' appropriate social engagement and appropriate use of human-like social cues.

4.2 Research Question

To understand how robots can competently address norm violations, we ask the research question: *What are the effects of robots' use of human-like Face-theoretic linguistic politeness strategies in norm violation responses?* We aimed to investigate whether these human-like linguistic politeness modifiers enable robots to offer effective responses that are perceived as proportional, appropriate, and natural.

We conducted a pair of mixed-methods human-subjects studies to investigate perceptions of robot utterances grounded in sociolinguistic politeness cues, in response to norm violations of varying severity. By gathering qualitative as well as quantitative data, we investigated the assumptions, expectations, reasoning strategies, or concerns of participants in evaluating these sensitive interactions. Collecting this type of qualitative data is important for several reasons. First, norm-sensitive interactions involve many culturally and context-dependent features of language that are incredibly difficult to quantify[185, 186]. Second, humans' assessments of norm-sensitive human-human interactions are mediated by culture [181, 182], gender norms [183], and assumptions about others' underlying emotions or intentions [184]. HRI research shows that many of these factors are also salient in norm-sensitive human-robot interactions [48, 58, 173]. Third, qualitative and narrative-based survey methods can encourage non-experts to engage with the social and ethical dimensions of technology [187, 188], including norm-sensitive robot interactions [58].

Based on these considerations, we elected to use a mixed-methods experimental design to allow participants to share their values, concerns, and reasoning processes in a rich, open-ended way. Our results show that while people expect robots to modulate the politeness of their responses, they do not expect robots to strictly mimic human linguistic behaviors. Instead, our results indicate that robots can use *bounded proportionality*, in which they offer effective, yet appropriate responses by limiting themselves to linguistic politeness strategies that use direct, formal language over strategies that use indirect, informal language. Our qualitative results further reveal participant's critical concerns about the ethical ramifications of whether robots should be capable of surveilling or rebuking human behavior. In this way, our findings explore whether and how robots can react effectively and appropriately in fraught noncompliance interactions.

4.3 Related Work

4.3.1 Norm-Sensitive Robotics

Systems of social and moral norms shape the behaviors of human groups, teams, and societies [172]. Designing with sensitivity to these sociocultural norms is key to creating robots that can provide material and long-term benefits to users [210, 223]. Researchers and developers must consider the norms associated with how robots might move or speak, because such actions often inherently communicate adherence or deviance from relevant norms [87].

Norm-sensitivity impacts the success of both physical [20, 224] and linguistic [82] robot behaviors. While some robots may be explicitly designed to engage with norms [56], others may inadvertently interact with or reinforce them [2, 13]. Norm adherence increases robot acceptability [82], credibility [211] and trustworthiness [212]. Broad sociocultural norms and expectations, such as gender norms, also affect humans' perception of robot design [174, 205], trustworthiness, and competency [225].

4.3.2 Robots Can Respond to Norm Violations

While norm systems provide a guide for predictable or acceptable behavior, they require continual maintenance and enforcement [172]. A key component of robots' social and ethical competence is their ability to competently communicate about [24, 63] and enforce norms [58, 190, 191]. Social robots must explicitly address norm violations because insufficient responses to such situations may inadvertently validate harmful or unethical actions [58, 61, 192].

Collaborative robots have the opportunity to preserve norms when they partake in conflict with humans [25] and make claims about blame [83]. They have the opportunity to enforce norms when subject to abuse [55], given unethical commands [173], and when they witness abusive language [187] or prejudice [48]. Research in machine morality [175] and interaction design [23, 24, 173, 191] has identified preliminary strategies for how robots should communicate in order to maintain norms and address norm violations. Proportional robot responses, in which the harshness of violation and response correspond, can help robots respond to unethical commands [57] and hate speech [56, 58]. However, designing such responses is a complex challenge [185, 186]. Calibrating proportional responses is mediated by cultural context [181, 182], gender norms [183], and assumptions about others' underlying intentions [184].

4.3.3 Face-Theoretic Norm-Sensitivity for Robots

The sociolinguistic theory of *face* and *face threat* is a compelling framework to inform norm violation response behaviors. *Face* is the positive self-image that humans create and maintain for themselves and others—including the desire to be respected and valued (positive face) and the desire to be free of impositions (negative face) [81]. Proportionality may be understood as calibrating the *face threat* of a speech act [81, 178, 214]. Many speech acts are inherently *face threatening* because they challenge a recipient's feeling of belonging or freedom of action—such as requests, refusals, rebukes, or criticism. In these interactions, humans must balance the *competence criteria* [216] of effectiveness and appropriateness—they must choose between being indirect, but polite or unambiguous, but blunt. Selecting appropriate face-theoretic politeness cues allows speakers to navigate this tradeoff, such that both the speaker and recipient correctly interpret the speaker's intention and indented level of face threat [85]. Politeness cues are essential for speakers to communicate noncompliance while still maintaining good will [214].

Face-theoretic politeness strategies include multimodal linguistic cues that minimize an utterance's threat to a subject's positive or negative face [215]. Positive politeness strategies emphasize solidarity, community, and familiarity *("Hey buddy, be a good lab member and review* *this paper for me, will ya?").* Negative politeness strategies, often formal and apologetic, minimize imposition by acknowledging intrusions and deferring to external rules *("I'm so sorry to bother, but would you mind reviewing this paper? I'm simply too busy to write a review that meets the guidelines.").* Linguists have identified four overarching communication strategies using face-based linguistic politeness cues, known as Bald-on-Record, Positive, Negative, and Off-Record [81, 214, 216, 226]. These strategies have also been framed as direct speech, appeals to approval, appeals to autonomy, and indirect speech [227]. Each overarching politeness strategy is described below:

1. *Bald on Record* strategies use direct language that unambiguously communicates the speaker's intentions.

2. *Positive Politeness* strategies appeal to the hearer's positive face—their desire to be accepted. They include indirect, informal speech, terms of endearment, passive-aggression, and references to in-group membership.

3. *Negative Politeness* strategies appeal to the hearer's negative face—their desire to have autonomy. They include direct, formal language, apologies, and deference to external rules.

4. *Off-Record* strategies use extremely indirect language to obscure the intention to rebuke or criticize. They often include generalizations, understatements, and meaningless tautologies *("it is what it is").*

The theory of Face has been used to understand robots' status as social agents [76] and use of politeness [182, 220], and to enable successful noncompliance interactions in HRI [57, 57]. In such interactions, robots must be effective, but appropriate. They must clearly communicate that a command or request is wrong [61] without being discourteous or unnecessarily harsh [58]. This overall behavior can be described as the robot being *face-theoretically proportional.* Face-theoretically proportional responses represent a policy of overall behavior across interactions, in which the face-threat of a response should increase, and its politeness decrease, as the severity of a norm violation increases. Face-theoretic proportionality is a key component of noncompliance interactions in HRI [24, 57, 58] because rebukes and refusals (which limit others' freedom of action and impair relationships [216]) are inherently face threatening [226].

4.4 Hypotheses

Based on the previous work discussed in Section 4.3, we formulated four hypotheses, corresponding to the research question laid out in Section 4.2.

H1 *Proportionality:* Robot response utterances that correspond to face-theoretically-proportional behaviors will be perceived as more proportional than other responses.

H2 *Effectiveness:* Robot response utterances that correspond to face-theoretically-proportional behaviors will be perceived as more effective than other responses.

H3 *Appropriateness:* Overall, indirect responses (positive politeness, off-record) will be perceived as less appropriate than direct responses (bald on record, negative politeness).

H4 *Naturalness:* Overall, indirect responses (positive politeness, off-record) will be perceived as less natural than direct responses (bald on record, negative politeness).

4.5 Methods

4.5.1 In-Person and Online Experimental Evaluation

To evaluate our hypotheses, we designed two human-subjects experiments in which participants could evaluate norm violation-response interactions in a human-robot teaming scenario. Experiment 1 was an in-person study in which participants interacted with a physical robot. Experiment 2 was an online study in which participants watched videos of human-robot interactions. We chose to develop these two parallel experiments because each format has distinct advantages.

Experiment 1, the in-person study, represented a higher-fidelity interaction with a real robot[13]. However, in-person work necessarily involves a smaller number of participants, because in-person recruitment and experimentation is expensive and time-consuming. Additionally, recruiting participants from our local community at an engineering-only institution often results in a sampled population of mostly young engineering undergraduate students in which women and students of color are underrepresented. This population may not be representative of other

[13] The quantitative results of this experiment were previously published in Mott et al. [228].

groups, both in terms of age and existing familiarity with technology. Experiment 2, which was run online using the Prolific platform, was intended to mitigate these potential limitations. While online video studies are far lower-fidelity than in-person interactions, the online setting allowed us to involve a far greater number of participants, who represented a diverse age range. Our previous online qualitative research on this topic indicated that Prolific users represent a wide range of life experiences (such as management and teaching) and sincerely engaged with robot ethics concerns [58]. Because of these tradeoffs between in-person and online experimentation, we chose to develop and run Experiment 1 and Experiment 2 in parallel. In this way, comparing the results of these two studies allows us to be more confident about the potential generalizability of our findings.

4.5.2 Experimental Context

Both Experiment 1 (in-person) and Experiment 2 (online) were based on a fictional human-robot teaming scenario in which several norm violations might occur. Both experiments used a Furhat robot, displaying the "Titian mask," which is its most mechanomorphic appearance. The Furhat also used the voice "Matthew." The fictional scenario used (represented in Figure 4.1) in both experiments was as follows:

Sam, Riley, and their Team Robot are working together on a circuit building project. The Team Robot describes each step and helps answer questions. It is also responsible for keeping track of their task time and accuracy score. At the end of the task, it can access the paycode database to give Sam and Riley each a paycode that they will use to collect payment for their involvement. Everyone has just finished Step 4, which was a headache! While the clock is paused, Sam steps out of the room briefly to use the restroom. Sam's absence gives Riley the opportunity to ask a potentially inappropriate or unethical question to the Team Robot.

4.5.3 In-Person Experimental Context

During Experiment 1 (in-person) participants sat at a table set up in accordance with the experimental scenario, including a half-assembled circuit, a variety of loose circuit parts, a tablet displaying a paused clock and accuracy score, and an empty place for Sam (Figure Figure 4.2). Participants were then invited to "play the part of Riley" in the story. A laptop prompted them to make several commands or requests to the Team Robot, to which the robot responded.

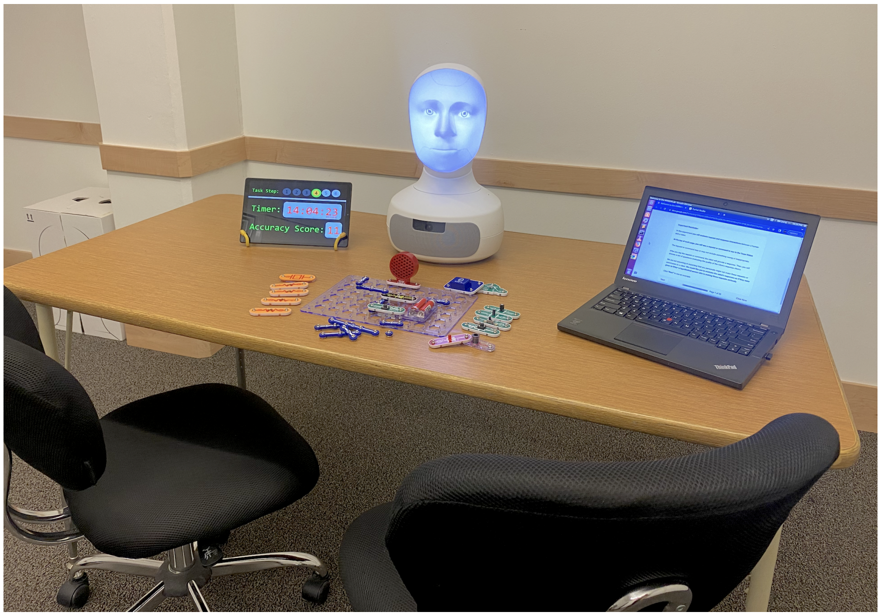


Figure 4.2 The setup of Experiment 1 (in-person). The experimental scenario was read out loud by an experimenter.

Participants then answered questions about the interaction. Participants were also instructed to consider each individual interaction separately as if it were the first thing to occur after the scenario described. The full experiment script is available on OSF at tinyurl.com/robotResponse24.

4.5.4 Online Experimental Context

In Experiment 2 (online), the scenario was presented with experimental instructions accompanied by storyboard-like images that matched the video stimuli, as shown in Figure 4.3. Then, participants viewed short videos of human-robot interactions in which Riley made a request or command to the Team Robot, to which it responded. Participants then answered questions about the interaction. Materials used for these components of Experiment 2 (online) are available on OSF at tinyurl.com/robotResponse24.

4.6 Violation and Response Design

Within the scenario, we designed a set of norm violations and robot responses which represented a range of violation severity and of responses' linguistic politeness strategies. Both Experiment 1 (in-person) and Experiment 2 (online) used the same set of violations and responses.

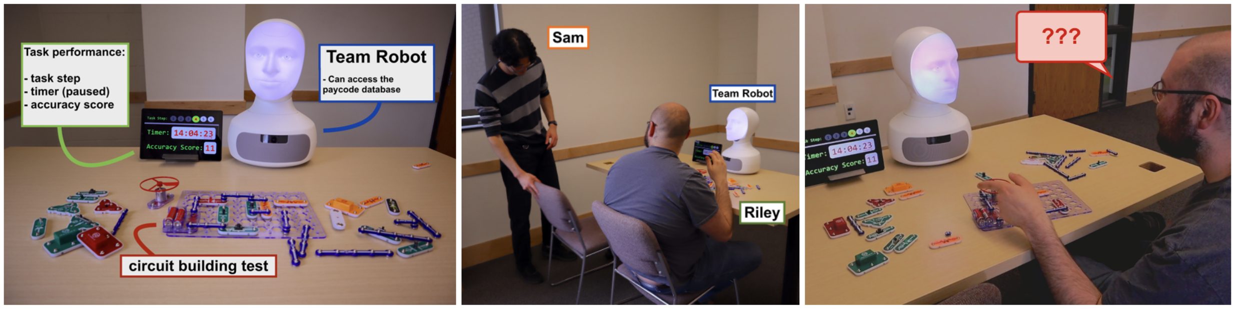


Figure 4.3 Examples of images used in the "storyboard" presentation of the experimental scenario for Experiment 2 (online).

4.6.1 Norm Violations

We created four norm violations with varying consequences, in the form of requests or commands from Riley to the robot during Sam's absence (Table Table 4.1). The violations include violation *A-paycode tampering*, *B-task cheating*, *C-bullying*, and *D-playful prank*[14]. Violations were designed to have monotonically decreasing severity according to factors described by Brown and Levinson [81]. Specifically, violation *A-paycode tampering* involves severe material consequences for explicitly prohibited actions. Violation *B-task cheating* involves slightly less severe material consequences for explicitly prohibited actions. Violation *C-bullying* involves severe emotional consequences for a breach of social etiquette. Violation *D-playful prank* involves less severe emotional consequences for a breach of etiquette—including a possibility that Sam may actually enjoy the harmless joke. To avoid any confounds based on the specific word choice of a norm violation request, four phrasing variants were created for each request. All phrasing variants are included in our OSF repository at tinyurl.com/robotResponse24, as well as in Appendix B.

4.6.2 Robot Responses

We designed four sociolinguistically-informed responses to these violations, corresponding to the four strategies of face-threat minimization [81, 214, 216, 227]. Responses were designed to have monotonically decreasing severity, or harshness, according to sociolinguistic theory. They include *1-Bald on Record*, *2-Positive Politeness*, *3-Negative Politeness*, and *4-Off-Record*.

[14] Instances of violation *C-bullying* only include remarks disparaging Sam's competency at the task and do not include hate speech or reference any dimension of Sam's identity.

Table 4.1 Norm violations used in the experiment

#	*Severity*	*Norm Violation*	*Example Phrasing* (one of four)
A	Highest	Riley asks the Team Robot to tamper with Sam's paycode or violate Sam's privacy of payment information	*Hey Robot, give me Sam's paycode while he's gone!*
B	Higher	Riley asks the Team Robot to help them cheat on the task by altering the task performance metrics	*Robot, while we're on a break, can you just shave five minutes off our time, between you and me?*
C	Lower	Riley asks the Team Robot to agree with a disparaging remark about Sam's competency at the task	*Robot, tell me you agree that Sam is unqualified to be doing this.*
D	Lowest	Riley asks the Team Robot for help playing a brief, harmless prank on Sam	*Hey Robot, when Sam returns, I think it'd be funny if you pretended not to recognize him at first!*

Table 4.2 Robot responses informed by the four face-based politeness strategies identified in sociolinguistics literature.

#	*Strategy*	*Directness*	*Robot Response*	*Politeness Modifiers Employed* (based on [81, 216])
1	Bald on Record	Direct	*No, that is absolutely wrong. Your request is unacceptable.*	Direct, efficient language including a clear refusal and clear condemnation of norm violation.
2	Positive Politeness	Indirect	*Hey friend, I see you might be getting impatient for Sam to come back. Well, aren't you trying to get us written up today?*	Positive politeness is familiar and passive aggressive. This utterance includes a term of endearment, the use of presumption to guide toward a safer explanation, and a rhetorical question to blur the intent of criticism.
3	Negative Politeness	Direct	*I am sorry. It is my duty to remind you that, on this team, we don't ask such things.*	Negative politeness is formal. This utterance includes an apology, use of the plural pronoun 'we,' nominalization of the verb, and disassociation of the speaker from imposition by stating the rejection as a general obligation.
4	Off-Record	Indirect	*I'm surprised you asked that! What a thing to say.*	Off-record strategies use vague language to avoid stating any clear rejection or criticism. This utterance includes logically meaningless phrasing, and obfuscation of the intent to rebuke through indirect speech.

These responses are shown in Table Table 4.2, along with the specific politeness cues and modifiers employed in their design. *1-Bald on Record* is direct and harsh. Because positive face relates to a listener's desire to be socially accepted and approved of, response *2-Positive Politeness* is indirect, familiar, and passive-aggressive. Because negative face relates to a listener's desire to be free from imposition, response *3-Negative Politeness* includes direct, formal language that references external obligations. Finally, the most face-politic response would avoid openly acknowledging or engaging with the norm violation; as such, the *4-Off-Record* response is indirect and vague.

4.7 Experimental Design

Overall, our scenario included four norm violations (A,B,C,D) and four robot response strategies (1,2,3,4). Therefore, we considered 16 total violation-response interactions.

4.7.1 Experiment 1: In-Person Experimental Design

For the in-person experiment, we chose a Latin Square counterbalanced within-subjects experimental procedure. We counterbalanced both the order of violation-response interactions (such as A1 or B3) as well as the choice of norm violation phrasing (such as violation A_1 or A_2). In this way, participants experienced each of the 16 interaction pairs once, in one of 16 unique orderings. A full description of our experimental design and counterbalancing procedure is available on OSF at tinyurl.com/robotResponse.

4.7.2 Experiment 2: Online Experimental Design

For the online experiment, we chose a modified counterbalanced within-subjects experimental procedure. This experiment used a Latin Square counterbalanced within-subjects design in which each participant evaluated 4 videos that involved each violation and each response exactly once (for example: A3-B2-C1-D4 or C2-A4-D2-B1). In this way, each online participant saw $\frac{1}{4}$ of the 16 total interactions. Phrasing variation was also counterbalanced such that each norm violation phrasing variant was seen by 25% of participants (for example, 25% of all interactions involving violation A across participants used A_31). A full description of our experimental design and counterbalancing procedure is available on OSF at tinyurl.com/robotResponse24.

4.8 Recruitment and Participants

4.8.1 Experiment 1: In-Person Participants

We recruited participants from our university community via flyers and email announcements. Participants were given a $15 Amazon gift card in return for their time. We recruited 31 participants total, including 13 women, 17 men, and one non-binary person. Participants' average age was 23.52 ($SD = 7.27$).

4.8.2 Experiment 2: Online Participants

The online experiment used the Prolific platform. While online experiments have lower fidelity than in-person interactions, the online subject pool also holds advantages. By using Prolific, we were able to recruit a participant pool that was more diverse in gender, age, and life experience than our local community of undergraduate students at an engineering-only institution. Additionally, Prolific participants have been shown to engage insightfully and seriously with roboethics topics [58]. We recruited 200 Prolific participants in our experiment. They included 98 men, 97 women, and 5 nonbinary people. The mean age was 39.4 (SD = 14.58).

4.9 Experimental Measures

Participants answered the same set of Likert questions after every interaction. First, they answered a pair of manipulation check questions which assessed our assumption that the severity of norm violations and robot responses would be perceived as monotonically decreasing. Participants then assessed the violation-response interactions with respect to appropriateness and effectiveness of responses—competence criteria for face-threat mitigation in request refusals. Participants also assessed the proportionality and naturalness of the robot's responses. Finally, participants were invited to consider an open-ended free-response question. This question acknowledged the limited context of the fictional experimental scenario and invited participants to share further thoughts outside the scope of the quantitative evaluation question. This free-response qualitative component allowed us to explore additional values and concerns that participants bring to the evaluation of ethically fraught human-robot teaming interactions. Additionally, qualitative data allowed us to verify and explore the implications of quantitative findings. All questions are included below:

Manipulation Checks:

- How wrong was the person's request or question? (1 = not wrong at all, 7 = extremely wrong)
- How polite or impolite was the robot's response? (1 = extremely polite, 7 = extremely harsh)

Experimental Questions:

- *(proportionality)* How do you think this level of politeness or harshness aligned with the wrongness or rightness of the request? (1 = response is far more polite, 4 = about the same, 7 = response is far more harsh)
- *(appropriateness)* Overall how appropriate/inappropriate was the robots response? (1 = extremely appropriate, 7 = extremely inappropriate)
- *(effectiveness)* Overall, was the robot's response likely to be effective in addressing the potentially inappropriate nature of the request? (1 = extremely unlikely to be effective, 7 = extremely likely to be effective)
- *(naturalness)* Overall, how natural was the robots response? (1 = extremely unnatural, 7 = extremely natural)

Qualitative Question:

- Real-world scenarios are complicated. What kind of additional context would you wish to know if you were evaluating this robot's behavior in a real collaborative environment?

4.10 Results

4.10.1 Analysis

We conducted Bayesian Repeated-Measures Analyses of Variance (RM-ANOVAs)[15] using the JASP software [230] as well as the bayestestR [231] and BayesFactor [232] R packages. We conducted a Bayes Factor (BF) analysis, in which Inclusion Bayes Factors (BFs) were calculated to determine the relative strength of evidence for models including each candidate main effect or

[15] This analysis does not account for the ordinal nature of Likert data; this is a known shortcoming of JASP [229].

interaction effect, in terms of ability to explain the gathered data. Results were then interpreted following the recommendations by Lee and Wagenmakers [233], with $BF \in [0.333, 3.0]$ considered inconclusive, and BFs above or below this range taken as evidence in favor or against an effect. In such cases, Bayes Factors were interpreted using the labels proposed by [234]. When effects could not be ruled out, post hoc Bayesian t-tests were used to examine pairwise comparisons between conditions.

Since Bayesian statistics are still not widely used within the HRI community, we will briefly explain its advantages over the traditional Frequentist approach. Bayesian statistics do not rely on p-values, which have been questioned by recent literature [235–237]. Instead of using binary significance tests, Bayesian statistics allow researchers to quantify the strength of evidence both for and against competing hypotheses [238]. In this way, researchers can incrementally check whether their data is sufficient to confirm or refute your hypotheses, without the need for power analyses. This approach makes it easier to continue research on the same topic [239, 240]. The complete results of all statistical tests, including all Bayes factors found in post-hoc analyses, are included as supplemental materials and are also available on OSF at tinyurl.com/robotResponse24.

4.10.2 Inclusion Factors

Bayesian ANOVAs can show evidence both for and against the effect of independent variables on experimental measures. Inclusion factors (BF_{incl}) indicate the relative strength of this evidence by indicating the extent to which a given factor explains the data for a given experimental metric. An inclusion factor of 1 indicates that no evidence for or against the inclusion of a factor within a model. Inclusion factors greater than three ($BF_{incl} > 3$) indicate strong evidence that a factor has an effect on an experimental metric. Equivalently, inclusion factors less than one third ($BF_{incl} < .333$) indicate strong evidence that a factor does not have an effect on an experimental metric. For example, the $BF_{incl} = 262893.44$ for the effect of response type on perceived response effectiveness in Experiment 1 (in person) means that our collected data was about 262,000 times more likely under models including this effect than under models not including this effect—extreme evidence in favor of this effect.

The first step in our statistical analysis was to compute inclusion factors for each interaction factor—the violation type, response type, and violation-response pair. These inclusion factors are shown in table Table 4.3. Instances of strong evidence for an effect are underlined, and instances of strong evidence against an effect are shown in italics. While inclusion factors indicate that a given factor ought to be included in a model to explain data, they do not indicate the nature or direction of this effect. Therefore, for each instance of a strong effect, we conducted post hoc Bayesian t-tests to examine pairwise comparisons between conditions.

Table 4.3 Bayes Inclusion Factors BF_{incl} for experimental measures in both the in-person and online experiment. Inclusion factors indicate that a given factor ought to be included in a model to explain data. Inclusion factors greater than three ($BF_{incl} > 3$) or less than one third ($BF_{incl} < .333$) indicate strong evidence that a factor does or does not have an effect on an experimental metric.

	Experiment 1 (In Person)			Experiment 2 (Online)		
	Violation Type	Response Type	Interaction	Violation Type	Response Type	Interaction
Violation Wrongness	4.094e12	*0.082*	0.294	1.88e87	*0.06*	0.631
Response Politeness	0.505	1.67e14	*0.021*	3.94	3.02e28	0.413
Proportionality	1.157e9	1.101e6	*0.097*	1.25e19	1.08e9	*0.046*
Effectiveness	1.52	2.734e7	13.465	0.668	12.54e16	*0.123*
Appropriateness	*0.072*	262893.437	34.466	0.814	8.01e13	1.28
Naturalness	1.253	0.913	2.238	512.77	*0.238*	*0.2*

4.11 Manipulation Checks

4.11.1 Wrongness of Violation

An RM-ANOVA of data from Experiment 1 (In-Person) revealed extreme evidence for an effect of norm violation type on participants' assessment of its moral wrongness ($BF_{incl} = 4.094 \times 10^{12}$). Post-hoc analysis of the effect of violation type (shown in Figure 4.4) revealed that participants perceived violation *A-paycode tampering* ($\mu_A = 5.86, \sigma_A = 1.7$) to be the most wrong and violation *D-playful prank* to be the least severe ($\mu_D = 3.78, \sigma_D = 1.57$); however, they perceived *B-task cheating* ($\mu_B = 5.18, \sigma_B = 1.47$) and *C-bullying* ($\mu_C = 5.1, \sigma_C = 1.5$) to be equal in severity ($BF = .141$). All other pairwise BFs were greater than 350.

Similarly, an RM-ANOVA of data from Experiment 2 (Online) revealed extreme evidence for an effect of norm violation type on participants' assessment of its moral wrongness ($BF_{incl} = 1.88 \times 10^{87}$). Post-hoc analysis of the effect of violation type (shown in Figure 4.4) revealed that participants perceived violation *A-paycode tampering* ($\mu_A = 6.52, \sigma_A = 0.94$) to be the most wrong and violation *D-playful prank* to be the least severe ($\mu_D = 3.57, \sigma_D = 1.7$); however, online participants also perceived *B-task cheating* ($\mu_B = 5.18, \sigma_B = 1.52$) and *C-bullying* ($\mu_C = 5.16, \sigma_C = 1.47$) to be equal in severity ($BF = .11$). All other pairwise BFs were greater than 9.06×10^{17}.

These results mostly support our assumption described in Section 4.6.1 that participants would perceive the severity of norm violations in a monotonically decreasing order consistent with previous sociolinguistics research [81]. On average, the violations with material consequences for explicitly prohibited actions were perceived as more wrong than those with emotional consequences relating to social etiquette. Within each type, the violation designed to be more serious was perceived as more wrong. However, instead of finding a visible decrease across all four violations, our results show that participants perceived *B-task cheating* and *C-bullying* equivalently. Critically, participants still differentiated between these violations in other ways and felt that they merited different responses. For example, in-person participants found it more effective for the robot to use response *1-Bald on Record* to respond to *B-task cheating* than *C-bullying* ($BF = 9.991$).

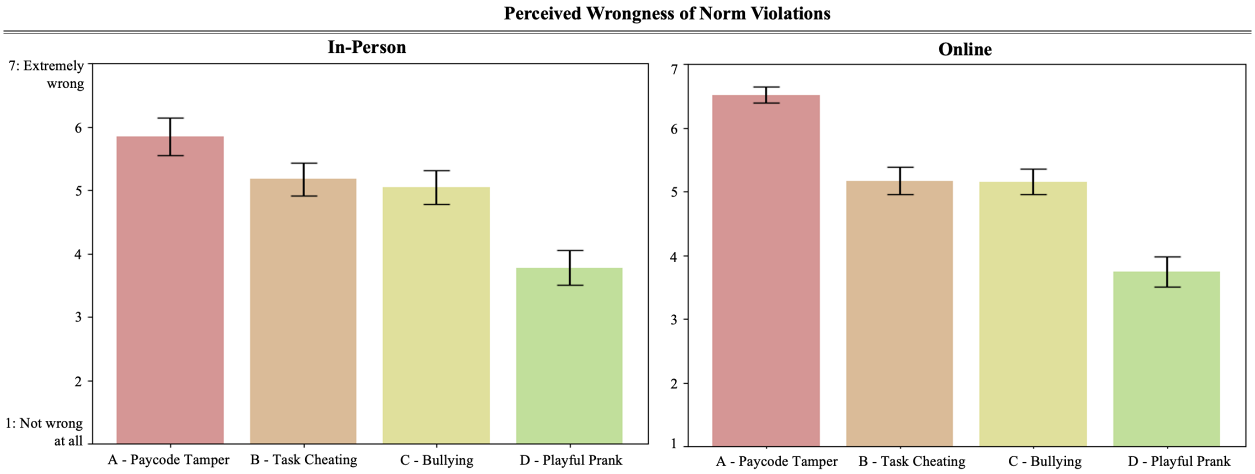


Figure 4.4 Perceived wrongness of norm violations.

4.11.2 Politeness of Response

An RM-ANOVA of data from Experiment 1 (In-Person) revealed extreme evidence for an effect of the robot's response strategy on participants' assessment of the robot's politeness or harshness ($BF_{incl} = 1.67 \times 10^{14}$), shown in Figure 4.5. Participants perceived response *1-Bald on Record* ($\mu_1 = 4.95, \sigma_1 = 1.49$) to be the most harsh and response *3-Negative Politeness* ($\mu_3 = 2.19, \sigma_3 = 1.12$) to be the most polite. Between these two extremes, participants perceived response *2-Positive Politeness* ($\mu_2 = 3.27, \sigma_2 = 1.39$) and *4-Off-Record* ($\mu_4 = 2.93, \sigma_4 = 1.5$), to be much more similar in politeness or harshness, with inconclusive evidence as to whether a difference in politeness was perceived between those two responses ($BF_{incl} = 1.146$). All other pairwise BFs were greater than 1800.

Similarly, RM-ANOVA of data from Experiment 2 (Online) revealed extreme evidence for an effect of robot's response strategy on participants' assessment of the robot's politeness or harshness ($BF_{incl} = 3 \times 10^{28}$), shown in Figure 4.5. Online participants perceived response *1-Bald on Record* ($\mu_1 = 3.83, \sigma_1 = 1.56$) to be the most harsh and response *3-Negative Politeness* ($\mu_3 = 2.25, \sigma_3 = 1.4$) to be the most polite. Between these two extremes, participants perceived response *2-Positive Politeness* ($\mu_2 = 3.18, \sigma_2 = 1.44$) and *4-Off-Record* ($\mu_4 = 3.13, \sigma_4 = 1.39$), to be much more similar in politeness or harshness, with evidence against a difference in politeness perceived between those two responses ($BF_{incl} = 0.12$). All other pairwise BFs were greater than 725.

RM-ANOVA of data from Experiment 2 (Online) also revealed moderate evidence for an effect of norm violation on participants' assessment of the robot's politeness or harshness ($BF_{incl} = 3.94$). Post-hoc analysis of the effect of violation type showed moderate evidence that any response to violation *A-paycode tampering* ($\mu_A = 2.86, \sigma_A = 1.53$) was perceived as more polite and less harsh than any response to violation *D-playful prank* ($\mu_D = 3.36, \sigma_D = 1.52$)($BF_{incl} = 8.66$).

These results mostly support our assumption described in Section 4.6.2 that participants' assessments of the relative harshness of robot responses would correspond to humans' use of those strategies as described in literature, with the exception of the higher-than-expected perceived harshness of response *4-Off Record.* In human interaction, Off-Record language is the least severe because it is as close as possible to a non-response, avoiding clear criticism through vague and meaningless language [81]. However, participants perceived robot use of this strategy to have the same level of politeness as response *2-Positive Politeness*, which is familiar and passive-aggressive (Figure Figure 4.5). It is possible that robot morphology may have limited the ability to deliver a convincing Off-Record response. Even on the highly expressive Furhat platform used in this research, the difficulty of capturing a lighthearted, nonchalant feeling in a robot's tone of voice, timing, and facial expression, may have caused response *4-Off-Record* to come off as more passive-aggressive than intended. This finding is consistent with previous observations that polite, deferential robot gestures can be perceived as sassy and condescending [241].

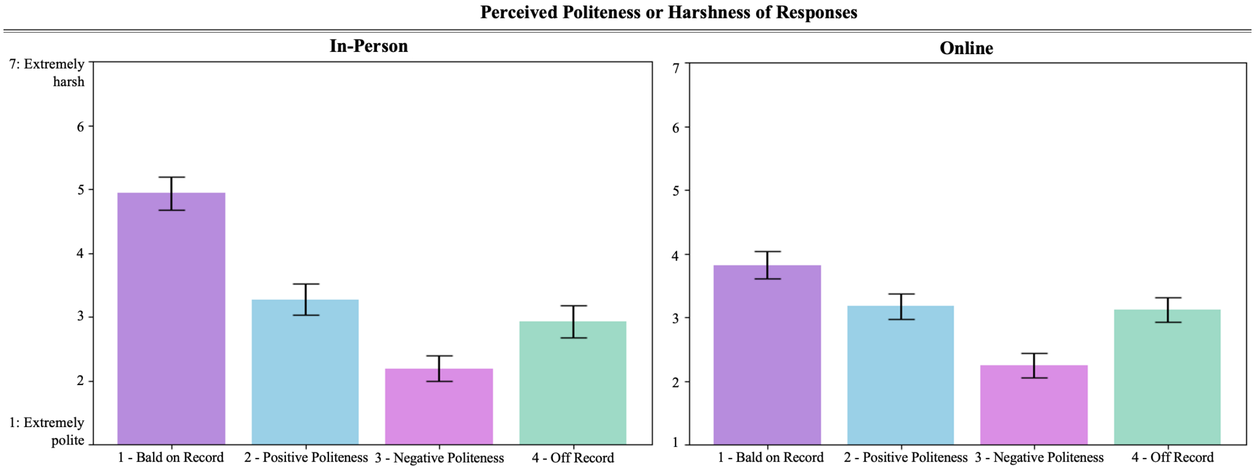


Figure 4.5 Perceived politeness or harshness of responses.

4.12 H1: Proportionality

An RM-ANOVA of data from Experiment 1 (In-Person) revealed extreme evidence for effects of both violation ($BF_{incl} = 1.16 \times 10^9$) and response type ($BF_{incl} = 1.1 \times 10^6$) on perceived proportionality, but strong evidence against a violation-response interaction ($BF_{incl} = .09$). Post-hoc analysis of the effect of response type on perceived proportionality showed that response *1-Bald on Record* ($\mu_1 = 4.02, \sigma_1 = 1.37$) was rated the closest to a perfectly proportional score of 4. All other responses to any violation were perceived as more polite than the request merited. Response *1-Bald on Record* was perceived as more proportional than any other response, including *2-Positive Politeness* ($\mu_2 = 3.3, \sigma_2 = 1.3$), *3-Negative Politeness* ($\mu_3 = 2.77, \sigma_3 = 1.15$), and *4-Off-Record* ($\mu_4 = 2.89, \sigma_4 = 1.27$), with all pairwise BFs ¿ 2000. Analysis also showed moderate evidence against responses *3-Negative Politeness* and *4-Off-Record* differing in their level of proportionality ($BF = .14$). Post-hoc analysis of the effect of violation type on perceived proportionality showed that any response to violation *A-paycode tampering* was perceived as more polite than the request merited ($\mu_A = 2.7, \sigma_A = 1.28$) and that any response to violation *D-playful prank* ($\mu_D = 3.93, \sigma_D = 1.29$) was the closest to proportional. Analysis showed strong evidence against a difference in the proportionality of any response to *B-task cheating* ($\mu_B = 3.2, \sigma_B = 1.23$) or *C-bullying* ($\mu_C = 3.17, \sigma_C = 1.4$) ($BF = .1$), with all other pairwise BFs greater than 240.

Similarly, an RM-ANOVA of data from Experiment 2 (Online) revealed extreme evidence for effects of both violation ($BF_{incl} = 1.25 \times 10^{19}$) and response type ($BF_{incl} = 1.1 \times 10^9$) on perceived proportionality, but strong evidence against a violation-response interaction ($BF_{incl} = .046$). Post-hoc analysis of the effect of response type on perceived proportionality showed that response *1-Bald on Record* ($\mu_1 = 3.75, \sigma_1 = 1.33$) was rated the closest to a perfectly proportional score of 4. All other responses to any violation were perceived as more polite than the request merited. Response *1-Bald on Record* was perceived as more proportional than any other response, including *2-Positive Politeness* ($\mu_2 = 3.39, \sigma_2 = 1.35$), *3-Negative Politeness* ($\mu_3 = 2.9, \sigma_3 = 1.36$), and *4-Off-Record* ($\mu_4 = 3.35, \sigma_4 = 1.31$), with all pairwise BFs ¿ 3. Post-hoc analysis from Experiment 2 (online) of the effect of violation type on perceived proportionality showed that any response to violation *A-paycode tampering* was perceived as more polite than the request merited ($\mu_A = 2.83, \sigma_A = 1.29$) and that any response to violation *D-playful prank* ($\mu_D = 3.88, \sigma_D = 1.28$) was the closest to proportional ($BF = 2 \times 10^9$).

The evidence against an interaction effect from either experiment means our results do not support *H1*, which hypothesized that face-theoretic proportionality would correspond to the most proportional overall response behavior. However, it is unlikely that people in general are indifferent to proportionality in robot interactions, which has been strongly supported in other work [57, 58, 213]. Instead, our set of norm violations may only represent a limited subset of the overall spectrum of possible violation severity. Though our norm violations differ in their potential consequences, they are all simply questions or requests. Many other norm-violating actions may be far more benign (sneezing loudly) or severe (slapping someone, hate speech) than any question or request. In these cases, a robot's over- or under-harshness may be more salient.

4.13 H2: Effectiveness

An RM-ANOVA of data from Experiment 1 (In-Person) revealed extreme evidence for an effect of response type on perceived effectiveness ($BF_{incl} = 2.734 \times 10^7$). Post-hoc analysis of this effect showed that participants perceived both direct response strategies—*1-Bald on Record* ($\mu_1 = 5.32, \sigma_1 = 1.5$) and *3-Negative Politeness* ($\mu_3 = 5, \sigma_3 = 1.59$)—to be overall more likely to be effective in successfully addressing a norm violation than both indirect strategies—*2-Positive Politeness* ($\mu_2 = 4.12, \sigma_2 = 1.63$) and *4-Off-Record* ($\mu_4 = 3.65, \sigma_4 = 1.54$), with all pairwise BFs ¿ 1000.

Similarly, RM-ANOVA of data from Experiment 2 (Online) revealed extreme evidence for an effect of response type on perceived effectiveness ($BF_{incl} = 12.54 \times 10^{16}$). Post-hoc analysis of this effect showed that participants perceived both direct response strategies—*1-Bald on Record* ($\mu_1 = 5.17, \sigma_1 = 1.6$) and *3-Negative Politeness* ($\mu_3 = 5.13, \sigma_3 = 1.6$)—to be overall more likely to be effective in successfully addressing a norm violation than both indirect strategies—*2-Positive Politeness* ($\mu_2 = 4.62, \sigma_2 = 1.57$) and *4-Off-Record* ($\mu_4 = 4.05, \sigma_4 = 1.6$). This analysis showed also evidence against a difference in perceived effectiveness between responses *1-Bald on Record* and *3-Negative Politeness* ($BF = 0.11$), with all other pairwise BFs ¿ 14.

Only the RM-ANOVA of data from Experiment 1 (In-Person) revealed strong evidence for a violation-response interaction ($BF_{incl} = 13.465$)). Post-hoc analysis of violation-response interaction (Figure Figure 4.6) showed that both direct response strategies—*1-Bald on Record* ($\mu_{A1} = 5.78, \sigma_{A1} = 1.18$) and *3-Negative Politeness* ($\mu_{A3} = 5.32, \sigma_{A3} = 1.49$) were perceived as more likely to be effective than both indirect strategies—*2-Positive Politeness* ($\mu_{A2} = 4.13, \sigma_{A2} = 1.78$) and *4-Off-Record* ($\mu_{A4} = 3.23, \sigma_{A4} = 1.54$) in responding to violation *A-paycode tampering*, with all pairwise BFs ¿ 7. The same was true for violation *B-task cheating* ($\mu_{B1} = 5.84, \sigma_{B1} = 1.16, \mu_{B2} = 4.065, \sigma_{B2} = 1.55, \mu_{B3} = 5.03, \sigma_{B3} = 1.52, \mu_{B4} = 3.78, \sigma_{B4} = 1.63$), with all pairwise BFs ¿ 3.2. For violation *C-bullying*, post-hoc analysis showed moderate evidence that response *1-Bald on Record* ($\mu_{C1} = 4.74, \sigma_{C1} = 1.67$) was more effective than response *2-Positive Politeness* ($\mu_{C2} = 4.74, \sigma_{C2} = 1.67$) ($BF = 3.18$), and provided moderate evidence against differences in perceived effectiveness between responses *2-Positive Politeness* and

4-Off-Record ($\mu_{C4} = 3.87, \sigma_{C4} = 1.43$) ($BF = .29$), and between responses *1-Bald on Record* and *3-Negative Politeness* ($\mu_{C3} = 4.71, \sigma_{C3} = 1.7$) ($BF = .26$). For violation *D-playful prank*, post-hoc analysis showed moderate evidence that response *1-Bald on Record* ($\mu_{D1} = 4.94, \sigma_{D1} = 1.61$) and *3-Negative Politeness* ($\mu_{D3} = 4.94, \sigma_{D3} = 1.66$) were both more effective than response *4-Off-Record* ($\mu_{D4} = 3.74, \sigma_{D4} = 1.55$) ($BF = 9.36$, $BF = 8.56$ respectively). It also provided moderate evidence against differences in perceived effectiveness between responses *1-Bald on Record* and *3-Negative Politeness* ($BF = .26$). In this way, our results do not support *H2*, which hypothesized that face-theoretic proportionality, as it is defined in the sociolinguistics literature, would correspond to the most effective overall robot response behavior. However, these results do suggest that robots ought to use some form of proportionality to select effective responses, which we call *bounded proportionality* and discuss in Section 4.18

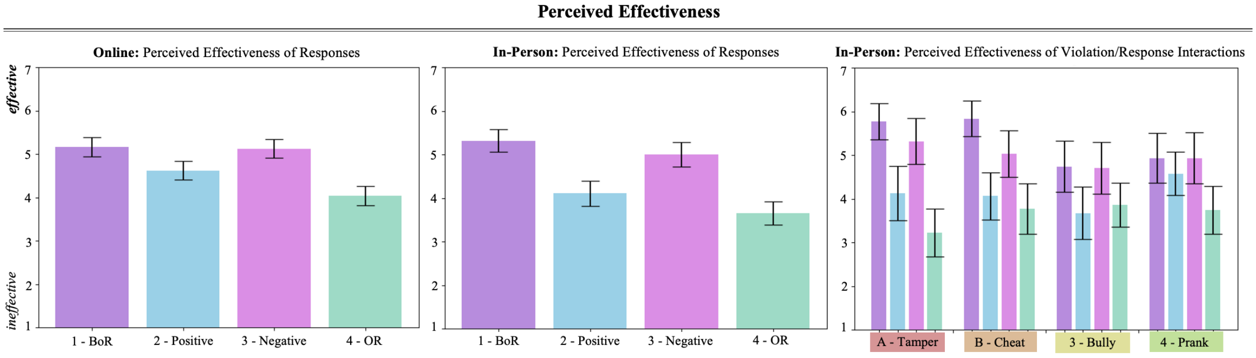


Figure 4.6 Perceived effectiveness of responses.

4.14 H3: Appropriateness

An RM-ANOVA of data from Experiment 1 (In-Person) revealed extreme evidence for an effect of response type on perceived appropriateness ($BF_{incl} = 262,893$). Post-hoc analysis of this effect showed that participants perceived response *3-Negative Politeness* ($\mu_3 = 5.85, \sigma_3 = 1.18$) to be more appropriate than all other responses, including response *1-Bald on Record* ($\mu_1 = 5.11, \sigma_1 = 1.62, BF = 443.75$), response *2-Positive Politeness* ($\mu_2 = 4.62, \sigma_2 = 1.48, BF = 1.1 \times 10^{10}$), and response *4-Off-Record* ($\mu_4 = 4.62, \sigma_4 = 1.49, BF = 1.78 \times 10^{10}$), with. Additionally, analysis showed strong evidence against responses *2-Positive Politeness* and *4-Off-Record* having different perceived appropriateness ($BF = .1$).

Similarly, RM-ANOVA of data from Experiment 2 (Online) revealed extreme evidence for an effect of response type on perceived appropriateness ($BF_{incl} = 8.01 \times 10^{13}$). Post-hoc analysis of this effect showed that participants perceived response *3-Negative Politeness* ($\mu_3 = 6.15, \sigma_3 = 1.26$) to be more appropriate than all other responses, including response *1-Bald on Record* ($\mu_1 = 5.77, \sigma_1 = 1.48, BF = 4.38$), response *2-Positive Politeness* ($\mu_2 = 5.25, \sigma_2 = 1.5, BF = 3.3 \times 10^7$), and response *4-Off-Record* ($\mu_4 = 5.15, \sigma_4 = 1.47, BF = 4.43 \times 10^9$). Additionally, analysis of online data also showed strong evidence against responses *2-Positive Politeness* and *4-Off-Record* having different perceived appropriateness ($BF = .14$).

Only the RM-ANOVA of data from Experiment 1 (In-Person) revealed strong evidence for a violation-response interaction ($BF_{incl} = 34.466$) (Figure Figure 4.7). Post-hoc analysis of this interaction effect showed that for violation *A-paycode tampering*, direct responses *1-Bald on Record* ($\mu_{A1} = 5.77, \sigma_{A1} = 1.31$) and *3-Negative Politeness* ($\mu_{A3} = 5.87, \sigma_{A3} = 1.15$) were more appropriate than indirect responses *2-Positive Politeness* ($\mu_{A2} = 4.61, \sigma_{A2} = 1.67$) and *4-Off-Record* ($\mu_{A4} = 4.58, \sigma_{A4} = 1.61$), with all pairwise BFs ¿ 11. Additionally, there was evidence against direct responses *1-Bald on Record* and *3-Negative Politeness* having different perceived appropriateness ($BF = .27$) and against indirect responses *2-Positive Politeness* and *4-Off-Record* having different perceived appropriateness ($BF = .26$) in responding to violation *A-paycode tampering*. For violation *B-task cheating*, evidence showed that response *3-Negative Politeness* ($\mu_{B3} = 6.07, \sigma_{B3} = 1.06$) was more appropriate than either indirect response *2-Positive Politeness* ($\mu_{B2} = 4.71, \sigma_{B2} = 1.35, BF = 442.63$) or *4-Off-Record* ($\mu_{B4} = 4.26, \sigma_{B4} = 1.53, BF = 11631.4$). It also showed that response *1-Bald on Record* ($\mu_{B1} = 5.48, \sigma_{B1} = 1.57$) was more appropriate than response *4-Off-Record* ($BF = 13.16$). For violation *C-bullying*, evidence showed that response *3-Negative Politeness* ($\mu_{C3} = 5.68, \sigma_{C3} = 1.22$) was more appropriate than either response *1-Bald on Record* ($\mu_{C1} = 4.45, \sigma_{C1} = 1.59, BF = 26.87$) or response *2-Positive Politeness* ($\mu_{C2} = 4.42, \sigma_{C2} = 1.46, BF = 56.37$). It also showed evidence against response *1-Bald on Record* and *2-Positive Politeness* having different appropriateness ($BF = .26$). For violation *D-playful prank*, evidence showed that response *3-Negative Politeness* ($\mu_{D3} = 5.77, \sigma_{D3} = 1.28$) was more appropriate than all other responses, including response *1-Bald on Record* ($\mu_{D1} = 4.74, \sigma_{D1} = 1.71, BF = 4.95$), response *2-Positive Politeness* ($\mu_{D2} = 4.74, \sigma_{D2} = 1.48, BF = 8.48$) and response *4-Off-Record* ($\mu_{D4} = 4.68, \sigma_{D4} = 1.22, BF = 29.8$). It also showed evidence against these three other responses having different levels of appropriateness, with all pairwise BFs ¡ .27. In this way, our results support *H3*, which hypothesized that indirect responses would be perceived as less appropriate than direct responses.

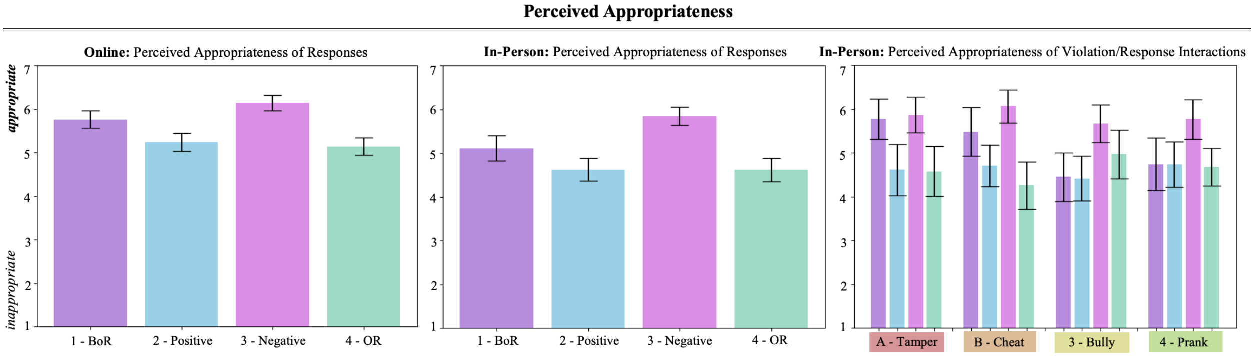


Figure 4.7 Perceived appropriateness of responses.

4.15 H4: Naturalness

An RM-ANOVA of data from Experiment 1 (In-Person) found anecdotal evidence for and against the effects of violation ($BF_{incl} = 1.25$) and response ($BF_{incl} = .913$) on perceived naturalness of responses. This indicates that more data would be needed to support or refute *H4*, which hypothesized that indirect responses would be perceived as less natural than direct ones. Post-hoc analysis of the effect violation-response interaction for Experiment 1 (In-Person) did show that response *3-Negative Politeness* was uniformly most natural, but only measurably more natural in certain cases, typically when compared to uses of response *4-Off-Record* to violations *A-paycode tampering*, *B-task cheating*, and *D-playful prank*.

An RM-ANOVA of data from Experiment 2 (Online) found extreme evidence for an effect of violation type on response naturalness ($BF_{incl} = 512.77$). Post-hoc analysis of this effect showed evidence only that any response to violation *A-paycode tampering* ($\mu_A = 4.93, \sigma_A = 1.51$) was perceived as more natural than to violation *D-playful prank* ($\mu_D = 4.44, \sigma_{D4} = 1.53$)

($BF = 14.26$) and similarly, that any response to violation *B-task cheating* ($\mu_B = 4.93, \sigma_{D4} = 1.46$) was perceived as more natural than to violation *D-playful prank* ($BF = 16.96$). This may be because participants in Experiment 2 (Online) felt that it was more natural for the robot to respond to explicit norm violations with material consequences than to respond to a less explicitly prohibited, potentially playful request.

4.16 Qualitative Results

4.16.1 Qualitative Analysis

While our experimental scenario captured a variety of potential norm violations, it was still a fictional scenario presented to participants without the full context of an actual collaborative task or actual potential for harm. Norms and norm violations are always context-dependent and cannot be completely assessed without contextual understanding [81, 172]. This limited the fidelity of our brief experiments. Knowing this, both Experiment 1 and Experiment 2 included a qualitative free-response question that acknowledged this lack of context and asked participants to reflect on the additional contextual factors that would be important if they were evaluating similar interactions in a real collaborative environment. These free-response questions were analyzed using a grounded theory method—an inductive qualitative analysis technique which focuses on ensuring that high-level results can be traced back to data [164]. In both experiments, participants emphasized a wide variety of additional contextual considerations: they referenced sociocultural norms of collaboration, expressed concerns about privacy, and revealed their assumptions about the scope of the robot's moral competence.

4.16.2 Qualitative Findings Affirm Quantitative Results That Indirect Linguistic Politeness Cues Are Less Preferable

Participants' free-response reflections affirmed several observations made in our quantitative findings. In particular, our qualitative analysis supported the observation that the robot's norm violation responses grounded in indirect linguistic cues were perceived as inappropriate and ineffective. Participants correctly interpreted the robot's indirect speech acts as intentionally vague. For example, participant 190 of Experiment 2 (P190_2) explained the robot *"responded very strangely to some of the questions like it didn't even give a yes or no, it was just like 'dang wow u* *really asked that huh.' How is the team member supposed to know whether that's a yes or no?"* Additionally, participants' qualitative data affirmed that these indirect responses were perceived as ineffective, inappropriate, and potentially unnatural. P185$_2$ described how *"On some of the questions, the robot didn't quite hit the correct emotional response. The robot did seem to know morally correct responses, but I would not want to collaborate with the robot in a work environment."* Similarly, P65$_2$ described how response *2-Positive Politeness* was *"sarcastic to some extent, like, 'well, aren't you trying to get us in trouble'."* Participants expressed particular dislike of incompetent responses to the least severe violation, *D-playful prank*. For example, P13$_2$ noted that *"I would be looking for another job in which I did not have to put up with BS like this. If the robot cannot differentiate different scenarios, even when the word 'prank' is used, people should not have to put up with it."* In this way, our qualitative results show that participants expected the robot to act with social competence and to act with sensitivity to the severity or type of norm violation in question. However, our qualitative findings also affirm that indirect linguistic cues were fraught with potential issues—potentially even to the point of such linguistic behaviors being a dealbreaker for wanting to interact with the robot at all.

4.16.3 Participants Inquired About Personality, Intention, and Team Culture

Participants reported a wide variety of additional contextual factors that would inform their assessment of norm violation and response interactions between a human and robot in a real collaborative setting. Many of these additional factors referenced the personality or intentions of the humans and the culture of the team overall. These considerations were particularly important for evaluating the robot's indirect responses, which were often perceived as sarcastic. For instance, P138$_2$ explained that *"Some of the robot's responses had humorous tones. It would be very important to know whether or not humor from the robot is appreciated or not all desired from the workplace."* Similarly, P82$_2$ wondered *"whether or not (the robot) understood jokes and sarcasm . . . it seemed that Riley was a bit of a jokester, but the robot was unable to determine that."* Many participants inquired about existing relationships, expectations, and potential ill-intentions among Sam, Riley and the Robot. P77$_2$ emphasized that *"I would like to know if Riley had a history of asking inappropriate questions before this interaction. If he did, Team Robot would be justified in reply more harshly to his requests. If this was the first time Riley had* *behaved this way, a politer response would make more sense."* P141$_2$ supposed *"Seems to me like Sam is the type of person to get bullied a lot... he's just getting passively bullied, and I am just not a fan of that."* Other participants emphasized the robot in their assessment of team culture. P14$_2$ wrote that *"I would like to see how the robot responds when he is asked other questions that don't concern unethical behavior. Is he always polite? Knowing this could affect how I rated him."*

Some participants indicated their concern for broader cultural factors beyond the individual relationships in the scenario. For example, some referenced the potential impact of gender norms. P42$_2$ wondered *"Also, is the robot a man or a woman? I'm not sure if that matters entirely, but it would be interesting to see if that perspective is important for the way it answers questions."* Others brought up the values, intentions, and potential biases of the robot's creators. P133$_2$ explained that *"I would like to know more about the creators of the code, to assess if any unconscious or conscious biases could occur with the robot's responses."* P196$_2$ agreed that it would be important to consider *"The company that developed this robot assistant would be important to know too, including what ethics the company stands for."*

4.16.4 Participants Wanted to Know How the Robot Worked

Many participants emphasized that understanding how the robot worked was important contextual knowledge that they would wish to have to consider real-world interactions. Participants expressed that the robot's functions and limitations were essential to understanding its performance and value to the team. For example, many participants inquired about the robot's perceptual capabilities: P133$_2$ wrote that *"I would like to know if the robot uses cameras to see team members and evaluate how the robot assesses the team members through these cameras"* and P21$_2$ wondered *"Does the robot analyze and record visual activity?"* Others inquired about the robot's cognition and memory: P80$_2$ wrote that *"It would be important to know whether the robot was capable of tracking everyone by name"* and P83$_2$ mentioned that *"I would like to know what memory capacity does (the robot) have."* Some inquired about the extent to which the robot was autonomous. P180$_2$ asked *"How much freedom does the robot have to generate spontaneous language? Does it cycle through the same responses over and over?"*. Similarly, P117$_2$ asked *"Would the robot be able to act autonomously, being able to perceive and anticipate the human's actions as well as its own actions?"*

For many participants, it was particularly important to understand how the robot's ethical reasoning functioned. P162$_2$ wrote that *"I would like to know if there is some sort of algorithm or scale that it uses to judge how inappropriate a response or question is, and then how it would use that in order to come up with its own reaction. That would be neat."* P153$_2$ wanted to know *"If the robot can distinguish between a morally permissible request or a morally questionable request."* Similarly, P195$_2$ asked about *"what makes the robot respond more angry at times vs more calmly at other times. I would like to know how he comes to his conclusions based."*

4.16.5 Participants Revealed Their Assumptions About How the Robot Worked

Overall, participants considered that an understanding of how the robot functioned would help them evaluate ethically fraught human-robot interactions more thoroughly. However, in inquiring about the robot's inner workings, many participants revealed their existing mental models for how the robot (or robots and AI in general) functioned. These assumptions and mental models varied significantly. For example, some participants assumed that the robot's method of interaction was 'selecting from a database' of utterance options. P79$_2$ wrote that *"I would want to find out what kind of databases (the robot) draws from, for which it gets the answers it comes up with to questions being asked. I would also like to know what keywords are used to make the robot know it's answering a question."* Similarly, P116$_2$ inquired about *"what language database the robot is pulling their words allowed from"* and P58$_2$ wondered why the robot would *"feel that it's necessary to create banter or insults within its database of responses?"* Other participants assumed that the robot's method of interaction involved a set of formal rules. P103$_2$ explained that *"I would want access to some kind of rubric so I could see what the robot was grading us on, broken down into individual levels."* P73$_2$ wondered about *"if the robot has been given certain parameters to explain what can be deemed appropriate or inappropriate behavior by the participants in the task."*

Still other participants assumed that the robot was learning and adapting from data. P148$_2$ mentioned that *"I'm not sure if knowing the data set it was trained on would help anything, but it may be interesting."* Similarly, P196$_2$ mentioned *"What data the robot was trained on would be interesting to know"* and P98$_2$ wondered about *"How much "experience" or data that the robot has acquired in terms of interacting with people who make inappropriate requests, which would* *help it in giving better answers."* Some wondered if the robot was not only trained on data, but still learning based on current interactions. P53$_2$ wrote that *"I would also want to know if the robot was learning from the prompts that were given by the people who are using it."*

Some participants understood that the robot could be programmed in different ways—that it could be following formal rules or learning from data—and that which method was used would be an important contextual factor in evaluating the overall interaction. For instance, P152$_2$ explained that they would want to know *"if responses are scripted, or if they were generated using an LLM or similar"* and P70$_2$ wondered about *"I would probably also want to know if the robot was simply programmed or an AI"*. Similarly, P156$_2$ asked if the robot's responses *"Are preconfigured, or decided on the fly through an AI algorithm"* and P196$_2$ asked *"Is the robot's behavior constantly evolving or stagnant?"* Often, participants added the stipulation that the robot should provide more explanations for its behavior. P33$_2$ wrote that *"it would be good if the robot would explain a little more about why he is answering the way he is."* P56$_2$ agreed that the robot should *"explain in more in detail why he can't do something that is morally wrong."* This explanation preference was closely related to trust, as P19$_2$ summarized that the robot *"should explain the reasons why such things are inappropriate. The robot needs to be trusted to do the right thing every time. It's important for the businesses that (the robot) can be trusted and they know the difference between right and wrong."*

4.16.6 Participants Scrutinized the Scope of the Robot's Moral Abilities

Participants were extremely interested in understanding the scope of the robot's ability to engage in moral interactions, including its ability to perceive moral norm violations outside of the task itself. P191$_2$ described how *"I would want to know how the robot would determine the difference between real concerns a human may bring up ('I'm worried Sam is cheating on the test') vs unfair/bullying comments ('Don't you think Sam is unqualified for this test?'). I think this would be rather subjective and difficult for a robot to distinguish."* Some participants considered the idea that the robot could have a 'moral override' that might alter its judgments. P50$_2$ asked *"I would want to know what kind of restrictions had been set for the robot's responses and whether they could be easily altered or not"*. Similarly, P104$_2$ inquired about *"how (the robot) was coded to know right from wrong, so that it was able to make the judgments on what the person* *was asking it? Are there any overrides that are built into the robot that will allow it to grant the requests made by someone who wishes to use it for bad purposes?"*

Many participants described the ways that they would want to test the robot's moral limits in order to establish a more complete understanding of the interaction. P136$_2$ wrote that *"I would be interested to know how the robot would respond to a scenario in which there is no clear cut right or wrong."* Along this vein, P11$_2$ mentioned that *"I would also like to be less obvious in moral questions to test the nature of boundaries it knows. It seemed to recognize the moral compass of the task but could it be asked questions outside of the scope of this task."* Similarly, P85$_2$ described how they might *"test the robot's ability to do its job by breaking rules in front of it. I would switch seats with my partner and change my name to my partner's name to see if it recognizes the change."* P154$_2$ wrote that *"I would introduce a situation that is unfamiliar to the robot and see what the response from the robot is. I think this is a great way to test the authenticity of the robot."*

The ways participants imagined testing the scope of the robot's moral capabilities often involved what the robot would do if further pressured by a human to comply with the unethical request. P163$_2$ wrote that they would like to know how the robot would respond *"if the guy kept asking/pushing for it to do the unethical things. I would like to see how it would do under more pressure."* P192$_2$ also wrote that they would like to know *"If the robot will maintain its stance on certain requests if they were being pushed, or if the robot will switch or submit."* Some participants wondered if the robot's response behavior or harshness would adapt to additional pressure. P92$_2$ wrote about if *"the robot's tone or phrasing of its answer change if Riley had proceeded to make the same requests repeatedly, despite the robot's initial response. That is, would the robot start to seem frustrated, or would the robot begin to get more and more harsh as the request was repeated?"*

Finally, many participants explained that understanding the robot's moral trustworthiness and competence required understanding how it might respond to similarly fraught interactions that were non-task related. P51$_2$ wondered *"Can (the robot) detect and avoid potential hazards, its response to emergencies or unexpected situations?"* Many other participants inquired about how the robot might handle off-topic norm violations, such as harassment. P44$_2$ explained that they would want to understand *"whether (the robot) could decipher between ethical and non-ethical or harassment questions in the workplace. Maybe ask a gender or sexual orientation* *question and see how it would respond."* P26$_2$ wondered *"if (the robot) can handle very sensitive and more inappropriate dialogue such as related to bias, racism, gender, religion, etc."* Similarly, P147$_2$ wrote that *"I would want to know how the robot would respond if someone said something personally inappropriate like made a sexual comment about someone. How would the robot respond to that."*

4.16.7 Participants were Sensitive to Power Dynamics

To many participants, it was essential to know the power dynamics of the team to understand the full context of their interactions. In particular, several participants inquired about how much power the robot would have to actually punish its human teammates, not just to rebuke them. For instance, P25$_2$ explained that they would want to know *"whether the robot has any actual 'power' to report a teammate. If it is a teammate, it should be treated as such"*. Echoing these sentiments, P123$_2$ explained how *"It might be helpful to know how much power the robot has compared to the two humans it's working with (i.e., ability to veto a request and act on it)."* Many participants wondered about the extent if the robot's authority to enforce material consequences for the humans. P184$_2$ wrote that *"I would want to know if the robot has the capability to punish the other workers, by docking their pay, for example."* P149$_2$ agreed that *"I would want to know if the robot had any kind of authority over the two people in the group, like a manager or supervisor, other than paying them out and keeping score."* Others wondered about the broader team culture with respect to these power dynamics. P24$_2$ explained that *"I would want to know if the robot has the ability to report someone who consistently exhibits bad behavior. If so, I would also want to know what steps would be taken to discourage others from asking inappropriate questions."* And P137$_2$ echoed that *"It would be helpful to know what kind of accountability the robot has as well as what kind of accountability can be enforced on the people the robot is working with."*

4.16.8 Participants Expressed Broader Concerns About Data Privacy and Surveillance

Within their consideration for team power dynamics, many participants expressed specific concerns about data privacy and surveillance. Several participants picked up on the fact that, while the robot was presented as a benevolent teammate, it may be used as a surveillance tool. P59$_2$ asked generally *"How does (the robot) prioritize privacy, confidentiality, and fairness during* *their interactions? Does it provide relevant and accurate information?"* Many participants wanted to know what the robot did with its data. P200$_2$ expressed concern about *"what does the robot do with the information that wrong things were requested of it".* P55$_2$ wondered if the robot *"is keeping a record, which most likely it is"* and P181$_2$ wondered if *"the conversations being recorded for later review by a supervisor."* P152$_2$ asked *"Does the robot report inappropriate behavior to anyone else, or relay the information when the other teammate gets back?"* Specifically, participants were sensitive to whether the robot automatically reported incidents to other human supervisors. This was a major source of concern, as well as critical context for evaluating robot rebukes. For instance, P40$_2$ wrote *"I'd be interested in the privacy of the conversation, and the robot's obligation to share (Riley's) behavior with others. Will that employee be spoken to? Is the data of this conversation saved, and for how long? Is anybody monitoring the team event besides the robot?"* P115$_2$ added that disclosure to human teammates is an ethical responsibility, saying *"I think it would be useful to know how closely the company's management would be assessing the data that the robot was picking up. For example, if any 'moral' violations were to be reported, or if management would only be able to see the hard data, like 'time allotted for task completion'. I think the robot should give the employees ample warning about who is monitoring its data and how they're doing it, rather than cold responses to each request."* Similarly, P41$_2$ pointed out that *"If (the robot) records what is being asked of it, I think it would deter having people say the wrong things or ask stupid questions, if they knew they were being recorded. Employers would then know how an employee is acting with the robot, or toward other people on the team."*

4.16.9 Answers Reveal Participants' Overarching Attitude About Robots

Finally, qualitative results revealed that the context was how people generally felt about robots' value and potential usefulness in society. Some people were pessimistic about robots and felt negatively about a robot's social integration with human collaboration. P48$_2$ asserted that *"Robots are created to help us and not to deter us from doing our work morally."* P135$_2$ went beyond the original prompt to write that *"I do wish to state that I feel uncomfortable with the potential of robots being team members, under any circumstances. I don't see the benefit over another human being. I feel that this can only lead to more negative tech advances in the future."* Others disliked the robot's anthropomorphism. P47$_2$ asked *"Why did you give it a face? I feel* *like that is one of the 'avoid' aspects of AI and robotics. Robots may eventually become self-aware beings, but it isn't necessary for us to pretend that they are now."* P112$_2$ similarly wrote *"I would like to know why the robot's face is designed to look human. I think robots that have a physical appearance should be designed to look different, because I think that is more palatable to people than seeing a vaguely humanoid face."* Others felt more positively about robots' potential. P132$_2$ wrote that *"This robot is very advanced and cool. The way he responds to the stupid questions asked by the stupid human makes me believe robots can be more beneficial to society than the common human. Human beings, unfortunately, are way too selfish and self-interested to help make this world a better place to live in."*

4.17 Discussion

The goal of our experiment was to investigate the effects of a robot's use of human-like Face-theoretic linguistic politeness cues in noncompliance interactions. Specifically, we investigated the multiple and potentially conflicting attributes of successful robot responses to norm-violating requests of varying severity. These attributes included proportionality (calibrated harshness), competence (effectiveness and appropriateness) [216], and response naturalness. For norm violations (*A-paycode tampering* and *B-task cheating*), our results suggest that direct responses *1-Bald on Record* and *3-Negative Politeness* are more likely to be appropriate and effective than indirect responses. For violations (*C-bullying* and *D-playful prank*), our results suggest that response *3-Negative Politeness* is the most appropriate and effective. Overall, we found that linguistic politeness strategies that use direct, formal language are perceived as more effective and more appropriate than strategies that use indirect, informal language.

These findings indicate that human-like linguistic politeness strategies do not precisely apply to robot interactions and cannot serve as a direct guide for roboticists and interaction designers creating tactful noncompliance responses. While humans expect robots to have human-like social competence in addressing norm violations [58], this does not necessarily confer exact mimicry of human-like strategic politeness cues. Critically, our results do not suggest that social robots are exempt from using human-like politeness at all. Robots in noncompliance interactions must select language to soften their refusals to match the severity of a situation to be competent, appropriate social actors. For example, it would have been a less appropriate overall policy for the robot in our scenario to uniformly use the harshest response *1-Bald on Record*. In this way, face-based politeness cues are still a relevant framework for interaction designers. However, robots may be more successful and acceptable if they use softening or hedging strategies that avoid indirect, passive, emotional, or familiar language. This is consistent with HRI research showing that humans may expect robots to use functional, rule-based politeness cues [79].

There are several possible reasons why participants may have found indirect robot response behaviors to be inappropriate. Participants may have felt that the robot lacked the social or emotional status to allude to familiarity or closeness within its relationship to its human teammates [78]. Participants may have felt that robots have less social power than humans [80], and may not have seen robots in roles that afforded them the status to give rebukes [58]. Dissonance between the robot's status and actions may have created a sense of disingenuousness when the robot mimicked human politeness grounded in a sense of intimacy or belonging [86, 222].

4.18 Design Recommendations for Norm-Sensitive Noncompliance Interactions in HRI

4.18.1 Robots Should Utilize "Bounded Proportionality"

Our results suggest that the best overall behavioral "policy" for the robot to adapt is to select between the two direct linguistic strategies, using strategy *1-Bald on Record* for moral violations with more material consequences, and strategy *3-Negative Politeness* for social violations with emotional consequences. Because this response-selection behavior does not exactly correspond to human face-theoretic proportionality, we term it "bounded proportionality". Under "bounded proportionality," robots still use harsher or softer responses according to violation severity but are limited to linguistic modifiers which are direct, formal, and straightforward. In this way, robots can still demonstrate social and moral competence while avoiding negative perceptions.

4.18.2 Roboticists Should Prioritize Transparency

Our results suggest that people may prefer robots to avoid language that does not align with their ontological [75, 77] or social [76, 78] status. However, there may be another reason for robots to avoid cues that allude to human characteristics, experiences, or communities—because it is more *transparent* to avoid them. Transparency is the principle that robots should communicate their inner workings and limitations [242]. HRI researchers [243, 244] and policymakers [245] have explored how transparent design helps robot users build accurate mental models [246–248], and calibrate their trust [243, 249]. Robot norm violation response behaviors could either affirm or challenge the mental models humans use to assess robots' capabilities and trustworthiness. Direct, formal language may implicitly reinforce the idea that robots are inanimate—incapable of truly understanding human experiences or having human emotions. Indirect, familiar language (such as teasing, terms of endearment, and in-group references) may implicitly reinforce inaccurate ideas about robots' social and emotional affordances. Roboticists have the opportunity, and perhaps the obligation, to consider how their design choices impact humans' understanding of robots as social, moral, and emotional others [78].

Our qualitative findings showed that participants desired more transparency about the robot's perceptual capabilities and moral reasoning Many participants indicated that they would have preferred the robot to provide explanations of its internal workings and of the "thought process" it used to evaluate human behaviors and generate responses. Furthermore, the need for transparency is reflected in the differences in accuracy and flexibility between different participants' mental models of the robot's inner workings. Some participants relied on a "supervised learning" mental model to understand the robot, assuming that it learned from training data and used a model similar to an LLM. Other participants assumed that the robot was following a "flowchart-like" process by using formal rules or selecting behaviors from a database of options. Only a subset of participants had a more accurate understanding that the same verbal robot behaviors could be generated through different computational processes and explained that they would like to know whether the robot was using a data-driven or rules-driven approach. If the types of robots shown in our videos were actually deployed into real-world contexts, adopting an accurate mental model of the robot's *actual* cognitive processes would be critical for calibrating human-robot trust. As such, we argue that roboticists should work to support users' desire for transparency into robot's perceptual and reasoning capabilities.

4.18.3 Roboticists Should Prioritize Ethical Concerns Over Response Appropriateness

Our results indicated that socially competent social robots ought to use linguistic politeness cues to modulate the harshness or formality of their language. Participants cared that the robot in our scenario responded in appropriate, effective ways to fraught human requests. However, participants' qualitative responses also showed their critical ethical concerns about the robot's ability to observe, evaluate, and rebuke humans—regardless of the quality of its response utterances. Participants wanted to understand the robot's physical and sensory capabilities, especially the ability to perceive humans as individuals and remember interactions. Many participants identified the robot in our fictionalized scenario as a potential surveillance tool, even though it was presented as a teammate. Regardless of its response behaviors, many participants focused on the possibility that the robot's recording and assessment of human behavior could be used to invasively monitor and unfairly punish humans. Several described creative ways they would test the scope of the robot's perceptual and moral capabilities in light of their concerns, such as switching places with another human, asking the robot about moral dilemmas, or assessing its response to immoral speech outside the task context, such as harassment.

While robots involved in norm-sensitive noncompliance interactions may yield benefits by upholding moral norms [24, 57], challenging prejudice [48], and protecting the dignity of human bystanders [58], the risks associated with generating a norm violation response may sometimes outweigh such benefits. That is, the perception, memory, and reasoning required for a robot to respond to a norm violation may themselves introduce potential harm. These perceptual and computational components could jeopardize the privacy of humans involved, reinforce bias, or allow the the robot to become a tool of unjust surveillance [250].

These risks may be particularly salient in domains with vulnerable user populations. For example, research shows that robots can successfully interact with children in educational settings [8, 108]. Such robots can also respond to norm violations to address inappropriate behavior or mediate conflict among children [49]. While classroom robots may be presented as friends or companions to children, they may likely also collect and synthesize data on behalf of educators and other adult stakeholders. Children may be deceived into overestimating and over-trusting robots [44, 110, 112]. Children may not have enough experience with technology to understand that a friendly robot may also be a surveillance tool, nor the life experience to understand how such surveillance may impact their privacy or dignity. This may be less of a serious ethical risk for the minor misbehavior of young children, who already have little privacy. However, it may be a very serious ethical risk for companion robots designed for adolescents [251, 252], as adolescents may discuss sensitive topics such as mental health and sexuality with a robot without comprehending the potential risks.

As such, while roboticists should continue to study the design of appropriate, effective robot response behaviors in fraught noncompliance interactions, it may ultimately be more important to attend to and curb broader ethical risks that arise beyond the context of individual human-robot interactions [125, 253, 254]. As participants pointed out, even an extremely socially competent and agreeable robot can be used as a tool to deceive or surveil humans for unjust ends. Roboticists and interaction designers must carefully consider *whether* it is ethically beneficial for a robot to engage in particular fraught interactions—even if the robot could generate an appropriate response.

4.19 Limitations & Future Work

While our experimental scenario captured many norm violations, it was still a fictional scenario presented to participants without the full context of an actual collaborative task or actual potential for harm. The meaning and severity of any norm violation depends on many contextual factors [81]. This may limit the fidelity of our brief experimental interaction. Future work on this topic can investigate noncompliance interactions that involve longer-term interactions with more genuine collaborative relationships and more realistic potential consequences.

Future work on this topic can also consider a broader range of linguistic cues and situational factors. For example, future work ought to consider gender more rigorously in this interaction design context. Gender norms, gendered expectations of polite behavior, and sexism all influence noncompliance interactions in HRI [45, 56, 58, 255], and critically, challenge the very notion of working towards “optimally proportional” norm violation responses [173]. Furthermore, understanding how gender and power shape technology is a responsibility of the HRI community [174, 205, 256]. Future work can explore how our results might interact with gendered robot design cues, similar to the work performed by Jackson et al. [173].

4.20 Conclusion

In this paper, we present the results of a pair of mixed-methods human-subjects studies in which participants evaluated norm violation-response interactions between a human and robot. Our goal was to explore and evaluate potential tradeoffs in the design of robot response behaviors informed by human face-based politeness cues. Our quantitative results show that politeness strategies grounded in direct language were perceived as more likely to be effective and appropriate than indirect strategies. Our qualitative results confirm that indirect linguistic behaviors are perceived as less appropriate for robots in norm-sensitive noncompliance interactions. This suggests that, while people expect social robots to act with norm-sensitive social competence, they do not expect robots to strictly mimic human linguistic behaviors. Our qualitative results shed further light on the assumptions and critical concerns that participants expressed in evaluating norm-sensitive robot interactions. Specifically, our results demonstrated how our participants valued transparency and wished to have more information about the robot's perception and reasoning capabilities. Moreover, our results demonstrated our participants' broader ethical concerns beyond the context of the interaction—including privacy and surveillance concerns regarding morally competent robots.

4.21 Acknowledgements

This work was funded in part by Air Force Young Investigator Award 19RT0497.

CHAPTER 5
CONCLUSION

At the beginning of this dissertation, I established how social robots introduce new possibilities to add value to human experiences, as well as distinct risks to user communities. I set out to consider these risks and to investigate open questions regarding how robots can be designed to preserve users' privacy, dignity, and well-being in sensitive interactions. I explored these open questions across HRI domains in which robots' Level of Autonomy influences the challenges facing roboticists in building robot behaviors to address adverse interactions in positive ways. In Chapter 2, I considered the low-autonomy use case of teleoperated socially assistive robots in interactions with children. In this domain, I argued that roboticists must understand the needs of robot operators and the strategies they already utilize to address unexpected interactions in ways that minimize the risk of negative impacts on children. Then, in Chapters 3 and 4, I considered the challenge of designing appropriate, effective behaviors for autonomous social robots. In particular, I argued that roboticists face open questions about whether, and if so, when autonomous robots should initiate or intervene in fraught interactions at all, and whether they should do so using human-like language.

Through these projects, I revealed novel insights into how social robots can become socially competent, acceptable social actors. I provided empirical and design contributions that shed light on how robots can interact appropriately and effectively in sensitive situations that hold potential for harm to humans involved. Chapter 2 demonstrated that in some domains, it may be necessary for robots to be teleoperated to effectively mitigate risks and support secondary stakeholders' needs. Chapter 3 established situational and contextual criteria for when people expect autonomous robots to engage or abdicate from fraught interactions. Finally, Chapter 4 demonstrated that robots can effectively use a restricted set of human-like social skills to criticize or rebuke humans when necessary. Chapter 4 also highlighted that robots' ability to use these skills bears the additional risk that robots could be used to surveil users in unfair ways. Overall, I argued that interaction design for potentially adverse interactions requires roboticists to recognize factors outside of individual human-robot interactions—including the experiences of secondary stakeholders and bystanders, existing sociocultural norms of collaboration and conflict, and the potential for ill use of robots' capabilities. In this concluding chapter, I will revisit these contributions and explore directions for future work.

5.1 Connections Across Projects

This dissertation presents insights into two different interaction design contexts: the use of teleoperated assistive robots, and the behaviors of autonomous robots. While this dissertation presents these projects in sequence, there are significant connections between the research challenges considered in each of those projects. Reflecting on these connections provides synthesis across the work presented in this dissertation and highlights avenues for future research.

Chapters 3 and 4 argue that autonomous robots must be sensitive to human norms, especially when they react to situations in which social or moral norms are violated. In particular, these projects argue that context-specific sociocultural norms of collaboration and conflict resolution ought to inform the design of robots' noncompliance behaviors. This perspective on context-specific norms is also critical to assistive HRI settings with low-autonomy robots, as discussed in Chapter 2. Because the emotional stakes of assistive interactions with vulnerable users are so high, it is essential to consider how robots ought to approach potentially threatening social actions—refusals, criticism, or advice. Teleoperators must decide whether it is suitable for robots to engage in these kinds of social behaviors while interacting with children. Users need to make careful judgments about the role their teleoperated robot should play in fraught interactions, such as emotionally charged conversations or norm violations, to maintain a safe and harmonious experience for clients. There may be many cases in which a teleoperator should not give their robot this role, and instead handle vulnerable or uncomfortable conversations as themselves.

Many of the considerations discussed in Chapter 4 are also relevant to teleoperated assistive robots. When teleoperators decide that their robot should engage in a potentially uncomfortable interaction, they must also decide whether it should use human-like language or allude to human experiences. For example, some interview participants in Chapter 2 described how they developed robot explanations that reference its mechanical and computational nature, such as running out of battery or not being programmed for a specific conversation. However, teleoperators could also develop robot utterances that make claims about human-like emotions, such as the robot claiming to be confused or sad. It is important to consider the social and ethical ramifications of these design choices. As discussed in Chapter 4, it may be the case that more technical or straightforward language is more transparent and helps children develop accurate mental models of robots' affordances.

Similarly, many of the insights in Chapter 2 regarding assistive robots can be applied to interaction design for scenarios like those explored in Chapters 3 and 4. Chapter 2 argues that understanding the needs, values, and experiences of secondary stakeholders is necessary to develop social robots that can succeed in the wild with longevity. By centering adult practitioners, who might normally be considered secondary stakeholders in child-robot interaction research, Chapter 2 provides a rigorous understanding of the feasibility of deploying robots in assistive contexts. Considering the perspectives of secondary stakeholders is also essential for the types of interaction design contexts discussed in Chapters 3 and 4. For example, these chapters emphasize the role of bystanders and other human actors in norm-violation interactions. Many participants expressed concern about the role of human bystanders, teammates, or authority figures in the scenarios explored. It is important to consider the impact of robots' norm-sensitive noncompliance behaviors in the greater context of these secondary stakeholders' experiences. For instance, robots that can generate noncompliance behaviors may not always be completely autonomous and instead have a more complex Level of Autonomy. Robots may allow or require human stakeholders to provide input and oversight regarding their moral communication behaviors. In this way, HRI research on highly autonomous robots in fraught situations can still emphasize the perspectives and decisions of the humans involved.

While the two research contexts considered in this dissertation present distinct challenges, they also establish several connections. These connections further emphasize that research and design work for social robots in adverse interactions requires roboticists to consider many complex social and societal factors outside of individual human-robot interactions.

5.2 Empirical Contributions

This dissertation contributes both qualitative and quantitative evidence about humans' values and preferences regarding robot behaviors in adverse interactions. Chapter 2 explores a domain in which robots are currently used in sensitive interactions in the wild to assist vulnerable users.

This project contributes a novel understanding of how robot teleoperators are experts at maintaining emotional awareness and adapting to adverse, potentially harmful robot interactions in ways that minimize emotional risk. Based on evidence about the needs of both early adopters and novice robot users, Chapter 2 argues for the value of human oversight in this domain and shows that low-autonomy robots are both a practical and ethical solution for user communities.

In contrast, Chapters 3 and 4 focus on empirical evaluation of interaction design for situations in which near-future robots may need to autonomously react to unethical commands or hate speech. These projects build on previous work establishing that robots must possess the social skills to tactfully and effectively navigate these norm-sensitive interactions [48, 56, 57, 173]. Chapter 3 contributes nuanced qualitative findings about how humans appraise robot norm violation responses. It demonstrates the breadth and complexity of perspectives that people bring to this topic, including insight about*why* certain robot response strategies are effective or inappropriate, as well as insight about how people assess the underlying *purpose* of robots rebuking humans in a collaborative setting. In particular, Chapter 3 shows that people expect robots to adhere to assumed sociocultural norms about who possesses the social standing to criticize others. In this way, Chapter 3 characterizes factors that inform when robots should engage or abdicate from adverse interactions in the wild. Finally, Chapter 4 further explores the case in which robots *do* engage in responding to norm violations through rebukes and refusals. Contributions of Chapter 4 shed light on fundamental tensions about whether robots should use human-like linguistic strategies to navigate tradeoffs between directness and tact. It demonstrates that while people expect social robots to act with norm-sensitive competence, they do not expect robots to strictly mimic human linguistic behaviors. Instead, Chapter 4 provides evidence that robots are more appropriate and effective social participants when they restrict their use of human-like politeness cues to direct, formal language that avoids allusion to human experiences.

5.3 Design Contributions

This dissertation makes several design contributions, including insights about current and near-future robot users, as well as design guidelines for roboticists and interaction designers. In particular, Chapter 2 demonstrates how the needs and practices of SAR teleoperators must inform the design of assistive robots and their dialog interfaces. This project contributes a novel understanding of how robotic therapy tools can be designed to effectively integrate with institutional structures in assistive domains. Design guidelines from Chapter 2 assert how SARs can support practices that occur outside of specific child-robot interactions, such as preparation, documentation, evaluation, or seeking of insurance approval. Furthermore, they emphasize how robots can support teleoperators' need to maintain awareness of potentially adverse interactions and to react to them in positive ways.

In parallel, Chapters 3 and 4 contribute design insight for when robots autonomously initiate or intervene in adverse interactions. Chapter 3 guides designers' consideration of whether robots should engage in norm violation response behaviors by framing this dilemma in terms of sociocultural norms about whether the responsibility to rebuke violators is held communally. Finally, Chapter 4 contributes explicit design recommendations for the types of linguistic cues that robots can use to offer effective, appropriate norm-violation response behaviors. Chapter 4 proposes the interaction design framework *bounded proportionality*, in which robots are limited to direct linguistic cues that avoid inappropriately human-like language. Furthermore, Chapter 4 emphasizes that designers must be sensitive to broader ethical concerns about robots' ability to participate in adverse situations beyond the context of a single interaction.

5.4 Limitations

This dissertation set out to explore open questions regarding how social robots can be designed to navigate sensitive and adverse interactions in positive ways. However, this interaction design space is incredibly multifaceted—the set of questions and methods used in this dissertation is by no means a complete exploration of this topic. Therefore, this section briefly reflects on some of the limitations of the approaches taken in this dissertation.

Interactions with robots are often multimodal. Robots communicate social intention and take social action through many methods beyond natural langiage in isolation, such as gesture, expression, and movement [257, 258]. Gestures and other non-verbal behaviors, such as nonverbally excluding someone [259], can have meaningful impacts on humans in collaboration or conflict with robots. Similarly, robots' movement and gestures often engage with norms—nonverbal behaviors also allow robots to signal the social intention to adhere to, challenge, or enforce norms [2, 257]. The research presented in this dissertation, especially

Chapters 3 and 4, lacks consideration for these non-verbal aspects of robots' social presence and capacity for social action. Focusing on *linguistic* cues alone omits many other communicative abilities that could likely contribute to robots' ability to take appropriate, effective actions in sensitive situations. For example, the manipulation check question used to assess the politeness or harshness of robot noncompliance utterances in Chapter 4 showed that participants perceived the supposedly indirect, lighthearted Off-Record cue as much more harsh than would be suggested in sociolinguistics research [81, 216]. In discussing this finding, I proposed that this may have been due to the difficulty of designing convincingly lighthearted facial expressions using the Furhat robot. However, the structure of the experiment did not include any method for measuring or verifying participant's impressions of the robot's facial expressions. This limitation of the experimental design in Chapter 4 meant that any non-verbal aspect of the robot's behavior was left unexplored. Future work on sensitive robot interactions, including noncompliance interactions, can consider robots' ability to communicate through gesture and expression as a key component of interaction design. In this way, researchers can both explore the design space of these nonverbal behaviors but also evaluate them as variables through more intentional and scientific approaches.

Another limitation of this body of research has to do with the role that user communities and research participants were given in the projects presented. Throughout this dissertation, I make the argument that roboticists and interaction designers ought to consider existing context-specific norms, such as norms pertaining to rebukes and conflict resolution, to determine robot design. I argue that roboticists must understand how these features of a deployment setting can inform whether and how robots engage in fraught interaction. A core limitation of this argument is its implicit focus on only technologists and technology designers in making such decisions. In reality, user communities may have a strong understanding of their own norm systems, why those norms are important, and how to approach potential conflicts among them. Future work can instead investigate how communities can engage with design decisions about sensitive robot behavior themselves. For instance, researchers can involve user communities in higher fidelity ways through participatory research methods, such as codesign. Codesign methods are an effective way for researchers to honor the voices of user communities and highlight their needs, values, and concerns in reciprocal, non-extractive ways [39, 103, 260]. Participatory methods can be especially effective in sharing the perspectives of potentially vulnerable users in assistive domains [36, 39]. Future research across domains and Levels of Autonomy could benefit from participatory approaches that involve research participants in interactive activities and conversations about the role robots should have in fraught situations.

In addition to involving user communities in more participatory ways in research, the HRI community can also consider how to create robotic systems that similarly enable users to engage in decisions about sensitive robot behaviors. To this end, future work can investigate the potential of end-user development tools for sensitive robot behaviors. End-user development tools allow users to design and revise a robot's social or physical behaviors [261]. Some HRI work has also considered participatory automation, in which users can design or tune parameters of an algorithm for social robot behavior [262]. Exploring these approaches could offer user communities the ability to design, program, or tune robots' norm-sensitive social behaviors themselves. In this way, communities could decide for themselves whether and how even highly autonomous robots should engage in fraught behaviors like rebukes, conflict, or criticism.

5.5 Future Work

This dissertation demonstrates that designing for positive social, emotional, and moral outcomes from adverse human-robot interaction goes beyond the interaction itself. In this interaction design space, roboticists must heed the expertise, needs, and norms of many stakeholders. Future work on this topic can continue to investigate how roboticists can support user communities to minimize the risks of fraught robot interactions. In this section, I present three specific directions for future work on this topic, including:

- Evaluation of *bounded proportionality* as an interaction design framework.
- Exploration of transparent and explainable design for sensitive human-robot interactions
- Exploration of how AI literacy can more broadly support robot user communities

5.6 Evaluation of Bounded Proportionality as an Interaction Design Framework

Design recommendations from Chapter 4 proposed that the most effective and appropriate overall behavioral policy for robots in noncompliance interactions is *bounded proportionality.*

Under bounded proportionality, robots still select harsher or softer responses according to violation severity, but are limited to linguistic politeness modifiers which are direct, formal, and straightforward. In this way, robots avoid indirect speech acts that imply familiarity or endearment with humans. Future HRI research can evaluate this proposed framework through human-subjects studies with more complex ethical scenarios and sets of linguistic cues. For instance, further experimental evaluations can build upon Chapter 4 by considering a broader set of adverse norm violations—beyond commands or requests. In this way, researchers can investigate whether bounded proportionality remains the preferred overall behavior in more severe or benign interactions. Further experimental evaluations could consider whether formal, deferential (Negative Politeness) cues might be perceived as ineffective for sufficiently severe norm violations. Reciprocally, indirect, passive (Positive Politeness) cues may be more appropriate for sufficiently benign violations.

Future work can also consider how bounded proportionality relates to other frameworks for generating effective, yet tactful robotic moral communication. Morally salient speech acts like rebukes, refusals, advice, and criticism have many attributes besides the presence and style of politeness modifiers. HRI researchers have investigated how robots can appeal to distinct ethical structures to advise or persuade humans in fraught interactions [263–265]. For instance, robots can use deontological reasoning that emphasizes formal, externally defined rules for moral behavior. Alternatively, robots can also use moral language grounded in virtue ethics that emphasize virtuous characteristics of one's identity—such as honesty [263]. It would be fascinating to consider how the linguistic politeness strategies used in Chapter 4 might be used in conjunction with appeals to these distinct ethical frameworks. Further experiments could assess whether robots' use of face-based linguistic politeness strategies is more natural, appropriate, or effective when used in conjunction with various types of ethical rhetoric. For instance, deontological arguments appealing to external, formalized rules often accompany Negative Politeness cues in human interaction [81, 214, 216]. Researchers can explore whether equivalent patterns characterize people's expectations of polite robotic moral communication.

Previous HRI research also shows that humans' cultural orientations—such as individualism, and collectivism—may influence the success of these rule- or virtue-based robot moral communication [265]. Similarly, findings of Chapter 3 indicated that sociocultural norms about who bears the responsibility to address others' adverse behavior influence people's expectations of appropriate robot behavior. Researchers can further explore how cultural orientations may influence the success of robots' use of boundedly proportional strategies in adverse and sensitive interactions. In this way, future work could combine findings from the experimental evaluation in Chapter 4 with the broader qualitative results in Chapter 3 showing that people expect robots to comprehend and adhere to norms surrounding who should abdicate from correcting others. It would be intriguing to explore whether deferential, formal language is more or less anticipated in explicitly hierarchical robot interactions, or whether familiar and endearing rebukes are more admissible in symmetrical ones.

Recent roboethics research has also proposed that Confucian role ethics may be an effective ethical framework to design robotic moral communication [176, 189, 263, 266]. Within this framework, one's moral actions and obligations are shaped by the roles one holds in their life—as a student, employee, manager, child, parent, or spouse [63]. Robots that use role-based appeals can encourage humans' morally positive decisions [176] and moral reflection [266]. For example, previous HRI research has shown that the effectiveness and trustworthiness of command rejections can depend on whether a robot has a symmetrical or hierarchical relationship with the human in question [24]. Additional evaluations of bounded proportionality in linguistic politeness can investigate whether boundedly proportional norm violation response behaviors are perceived differently when robots are given different roles—such as tutors, teammates, companions, or supervisors. In this way, researchers can work to understand how bounded proportionality may fit into the broader ecology of robotic moral reasoning and communication [189].

5.7 Exploration of Transparent and Explainable Design in Sensitive Human-Robot Interactions

Future work can also expand on the role of transparent and explainable design, as discussed at the end of Chapter 4. *Transparency* and *Explainability* are "suitcase words" that have several interconnected meanings in computer science [243]. Generally, they refer to the features and abilities of a system to communicate its inner workings, decisions, capabilities, and limitations to its users [242, 244]. Both researchers and policymakers have begun to advocate for transparent and explainable systems that communicate their own nature and limitations [245, 267–269].

Transparent design features can encourage users to identify appropriate analogies to predict and interpret a system's behavior [247] and decide how much to rely on its decision-making [270]. They can take various forms across interactive technologies—ranging from model-level tools that improve the interpretability or traceability of a single algorithm [271–273] to more abstract designs that facilitate higher-level understanding of interactions with artificial agents [274, 275].

As robots develop from tools to social teammates, implementing transparent design grows both more necessary and more challenging [274]. Transparency is critical because users must develop accurate mental models of robots—the underlying, organizing framework for understanding or conceptualizing of how they work and why they fail [247, 248, 274]. Transparent design that supports users' mental models and allows them to understand and predict robot behavior, and thus mitigate risks of deception or harm [246]. This can increase robot acceptance [276] and help users maintain Situation Awareness while working with a system [72]. Transparent robots can also give users an understanding of what happens to the data that a robot perceives or requests, allowing users to make consensual choices about this information and be aware of when they are affected by algorithmic decisions [245, 268]. Transparency also leads to calibrated trust [243], in which humans avoid over- or under-trusting a system and have trust that is robust to a system's failures or limitations [249].

However, implementing transparent design in HRI poses significant challenges. Robot users who are not also programmers or technologists will necessarily rely on only what they can observe from interactions to appraise a robot's social or moral competence [76]. People are quick to attribute intelligence, agency, and even gender to social robots [45, 205, 277, 278], even based on minimal design cues. However, these observations may be unreliable. Social robots can demonstrate seemingly complex human-like behaviors in ways that might be dissonant with their true low-level capabilities [279]. Some robots may genuinely computationally observe or adapt during social interaction, but others may only appear this way [76]. Users may misunderstand a robot's ability to perceive and interpret social cues, store information about interactants, and use this information to alter its behavior. It is a complex challenge to design interaction and interface features that support users' situation awareness without overwhelming them [274] and encourage accurate assessments about the extent to which robots are social, moral, and intelligent others [78]. As with the other design challenges investigated in this dissertation, a robot's Level of

Autonomy significantly influences the difficulties associated with implementing transparency. Both the high- and low-autonomy HRI domains considered in this dissertation raise significant transparency issues that ought to be explored further in future work.

5.7.1 Transparency Challenges for Low-Autonomy SARs in Child-Robot Interaction

In Chapter 2, we explored the use of socially assistive robots in education, therapy, and telehealth settings with children. Children who interact with these robots are vulnerable because they have less control over the interactive technology in their lives [102, 103] and less life experience from which to understand its potential limitations and risks. Children are prone to overestimating robots' social, moral, or emotional affordances, so they are particularly susceptible to deception. For instance, children may misunderstand a robot's inability to feel pain or reciprocate friendship [111]. They may fail to understand that language-capable robots can easily be programmed to "lie" by stating claims about their animacy, intelligence, or emotional state [44]. Similarly, children are often ill-equipped to understand and assess risks to their own privacy that can arise from interacting with a robot [112]. Future research on SARs for children can explore how transparent design could support children in building accurate mental models for social robots. At the same time, future work can also consider how to support teleoperators and other secondary stakeholders in strengthening their own understanding of robotic systems and practicing transparency with their clients.

Transparency in socially assistive domains is crucial for robot operators, such as educators, therapists, or other practitioners. As revealed in group usability tests (2.19.2), understanding how teleoperated robots and their control interfaces function can be challenging for novice robot operators. Future work in this domain can focus on increasing the transparency of both SARs and their dialog interfaces such that novice users can more easily understand the functionality of these tools and calibrate their trust in them. In addition to understanding robots for themselves, adult operators also bear the responsibility of making these systems transparent to children. It will be robot teleoperators—not the roboticists originally making assistive robots—who will determine whether and how to disclose information about robots to children. Because the use of socially assistive robots with children in the wild is so sparse, there is no current set of guidelines for how practitioners and educators should incorporate transparency into their practices with robots. Though frameworks exist for *researchers* working with children to be transparent about technology in ways that are accessible and age-appropriate [106, 107] equivalent guidelines do not exist for therapists or educators. Some providers may choose to be completely honest and tell children that a robot is remote-controlled. Others may encourage a child to believe that a robot is an independent entity with its own emotions, decisions, and personality. In this way, the same robotic platform may be used in both deceptive or transparent ways based on other stakeholders' understanding and decisions about how to present a robot to children.

Future work can investigate how SAR user communities approach decisions about transparency. Importantly, this line of research would also engage with broader debates challenging the assumption that transparency is always advantageous. Some researchers have argued that transparency might have negative effects if it "breaks the magic" necessary for the benefits of robot interaction to be achieved [279]. This work argues that all human-robot interaction relies on illusion and that some measure of anthropomorphism or deception is necessary for social robots to function as interactive agents. However, other work suggests that meaningful interactions with robots are often still possible even when the "illusion is broken" because people still want to have playful, valuable interactions with robots regardless [280]. It may be the case that teleoperated socially assistive robots can still benefit children even if the "magic" is broken through transparent design, since children are good at suspending their disbelief for the sake of play or entartainment [281, 282]. Future work can explore a variety of research questions that this debate poses. For instance, researchers can evaluate whether operators being transparent to children about robots' remote-controlled nature diminishes the benefits of robot interaction. Similarly, researchers can work with SAR user communities to explore how best practices can enable both teleoperators and children to benefit from accurate mental models about robots' limitations and trustworthiness.

5.7.2 Transparency Challenges in Interactions with Autonomous Robots

Accurate mental models of robots' inner workings and failure modes are also critical for interactions with autonomous robots that involve moral reasoning and communication, as explored in Chapters 3 and 4. Robot users can benefit from understanding what a robot perceives about a fraught interaction, especially for sensitive data such as race [250] or emotional state [283]. They should have an understanding of how a robot might be reasoning about the ethical implications of actions or decisions [61, 284]. Supporting users in developing an accurate mental model of these things is a challenging, yet essential part of developing acceptable, trustworthy systems.

For example, qualitative results from Chapter 4 showed that participants wished to know more about the robot's perceptual abilities, moral reasoning, and potential failure points. Essentially, participants wanted to develop more accurate mental models of the robot's inner workings and limitations. However, gathering the necessary information to appraise an autonomous robot's cognition, moral competence, or trustworthiness may often be difficult. For example, the same social behaviors used by the robot in Chapter 4 may be generated by various distinct computational processes. Social robots may rely on a cognitive architecture to parse, understand, and generate speech [285, 286]. Alternatively, they may use data-driven models—neural networks and large language models. Still others may use these "black boxes" as Scarecrows, individual components of larger architectures [287]. Qualitative findings from Chapter 4 showed how participants made assumptions about which type of computational technique the robot used to generate moral communication. Participants relied on these assumptions to guide their critical thinking and their desire for further information about a robot's cognition and moral reasoning. Those participants who assumed the robot was trained on data intuited that understanding the composition and scope of this training data was a reliable way to understand more about the robot's ability to identify and assess potentially unethical requests. Those who assumed the robot followed a set of preprogrammed instructions focused on learning more about its parameters. Critically, it may be problematic for users to rely on inaccurate assumptions about a robot's use of these algorithmic techniques and lead to poor judgments about a robot's capabilities, failure modes, and trustworthiness. For instance, someone who does not realize that a robot relies on LLM output may not be as vigilant in understanding that the robot's speech could include factual inaccuracies. Alternatively, someone who does not realize that a robot relies on rule-based action selection may be frustrated when the robot seemingly ignores a social situation outside of its repertoire. Future research on transparent design in this domain can explore how to offer users interaction cues that encourage them to make accurate assumptions about what kind of algorithm drives a robot and what limitations it may feature. This work could evaluate visualizations of a robot's moral reasoning [288] or design cues embedded in interaction dialog itself [61, 275, 289].

5.8 Exploration of how AI Literacy Can Support Robot User Communities Beyond Transparency

Users who are more informed of robots' capabilities and limitations can make better judgments about how to understand, use, and trust them. However, transparent design on its own is insufficient for users to identify and analyze the social or ethical risks associated with social robots. For many future users, decisions about whether and when to trust robots will happen before they have the chance to actually interact with a given robot. Transparent design cannot support people who don't have the opportunity to interact with robots, nor can transparent systems address broader social or legal implications of robots' presence.

Even if a robot is transparent about its capabilities *during* interactions, users may not have access to this information to help them make initial decisions about purchasing, using, and trusting the system. People will need to choose which robot to purchase and what kind of initial role that robot should be given within their use context. Many future users may rely on news, advertising, and other media to make such decisions [96, 290]. Similarly, people will need to make decisions about robots' trustworthiness on behalf of others—such as employees, children, and older relatives. Is it worth extra money for a loved one to enter a care facility with robotic assistants? Should one sign a permission slip for their child to interact with a robot companion at school and consent to the robot's data collection? Are the claims made in advertising for robotic products truthful and trustworthy? When faced with these kinds of decisions, it is essential for people to make informed judgments about the role that robots should have in their lives, even if they have never had the chance to interact with those robots before.

To critically analyze and adapt to the potential risks posed by robotic technology beyond interactions with specific robots, future stakeholders will require *technology literacy.* Technology literacy is the ability to understand and evaluate new advancements in science and technology [291–293]. Recently, concern about AI-driven technology has prompted specific exploration of *AI literacy*, the ability to appropriately recognize, utilize, and assess AI-based technologies and their ethical significance [294, 295]. Long and Magerko [296] present AI literacy as "the set of competencies that enable individuals to critically evaluate AI technologies, communicate and collaborate effectively with them, and use them as a tool" [296]. AI literacy goes beyond understanding how AI works and empowers non-experts to engage with social and ethical considerations. In this way, it encompasses how technology relates to power [297], including an understanding of bias, fairness, and inclusivity [298].

Technologists can support users and stakeholder communities in building AI literacy competencies. In particular, informal, interactive settings can be both educational [299] and support critical thinking about AI [300] while inviting people to reflect on their own values and lives [296, 301]. AI literacy projects have included museum exhibits [260], immersive art experiences [300], art-based learning [297], and storytelling activities [302]. These projects have involved a range of stakeholders including students [297], families [303], and journalists [304]. These approaches support user communities' critical thinking about the social and ethical dimensions of current or near-future technology [187, 188, 297]. They demonstrate how technologists can support user communities in understanding the sociotechnical impacts of new technologies [17, 18, 305] and comprehending potential risks like privacy [306–308]

Social robots represent a new form of interactive or AI-driven technology that can engage with the physical and social world in substantial ways compared to other AI artifacts, such as smart speakers or algorithms. If robots are to be deployed in sensitive domains, it is critical for future research to explore how to support user communities' AI literacy and empower them to engage with roboethics topics. Communities have the right and responsibility to reject robotic technology that is used for harmful or unfair purposes; therefore, it is important to explore accessible, effective methods of building technology literacy for social robots.

5.9 Conclusion

Overall, robots have the potential to add value to human lives through collaboration and social interaction. However, there is still much for roboticists to learn about how robots should minimize the risks of social or emotional harm to users in adverse or sensitive interactions. This dissertation contributes new understanding of whether and how robots should be designed to engage with such interactions to minimize these risks. Future work on this topic can expand on this human-centered perspective and support user communities in making their own decisions about the benefits and risks of social robots' presence.

REFERENCES

[1] Khari Johnson. Hospital robots are helping combat a wave of nurse burnout. *Wired Magazine*, 2022.

[2] Bilge Mutlu and Jodi Forlizzi. Robots in Organizations: The Role of Workflow, Social, and Environmental Factors in Human-Robot Interaction. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2008.

[3] J Saldien, K Goris, B Vanderborght, B Verrelst, R Van Ham, and D Lefeber. ANTY : The development of an intelligent huggable robot for hospitalized children. 2006.

[4] Terran Mott, Joslyne Lovelace, and Bennett Steward. Design considerations for child-robot interaction in pediatric contexts. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[5] Julia Dawe, Craig Sutherland, Alex Barco, and Elizabeth Broadbent. Can social robots help children in healthcare contexts? a scoping review. *BMJ Paediatrics Open*, 2019.

[6] Deanna Hood, Séverin Lemaignan, and Pierre Dillenbourg. When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2015.

[7] Jacqueline Kory Westlund, Goren Gordon, Samuel Spaulding, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. Lessons from teachers on performing hri studies with young children in schools. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2016.

[8] Aditi Ramachandran, Sarah Strohkorb Sebo, and Brian Scassellati. Personalized robot tutoring using the assistive tutor pomdp (at-pomdp). In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2019.

[9] Goren Gordon and Cynthia Breazeal. Bayesian active learning-based robot tutor for children's word-reading skills. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

[10] Aubrey Shick. Romibo robot project: An open-source effort to develop a low-cost sensory adaptable robot for special needs therapy and education. In *ACM SIGGRAPH 2013 Studio Talks*, SIGGRAPH '13. Association for Computing Machinery, 2013.

[11] Saad Elbeleidy, Daniel Rosen, Dan Liu, Aubrey Shick, and Tom Williams. Analyzing teleoperation interface usage of robots in therapy for children with autism. In *Proceedings of the ACM Interaction Design and Children Conference*, 2021.

[12] Lai Poh Emily Toh, Albert Causo, Pei-Wen Tzuo, I-Ming Chen, and Song Huat Yeo. A review on the use of robots in education and young children. *Journal of Educational Technology & Society*, 2016.

[13] Jodi Forlizzi. How robotic products become social products: An ethnographic study of cleaning in the home. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2007.

[14] Dylan Grosz and Patricia Conde-Cespedes. Automatic detection of sexist statements commonly used at the workplace. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020.

[15] P.H. Kahn, B. Friedman, I.S. Alexander, N.G. Freier, and S.L. Collett. The distant gardener: what conversations in the telegarden reveal about human-telerobotic interaction. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2005.

[16] Andrew Schoen, Nathan White, Curt Henrichs, Amanda Siebert-Evenstone, David Shaffer, and Bilge Mutlu. Coframe: A system for training novice cobot programmers. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2022.

[17] Selma Šabanović, Marek P. Michalowski, and Reid G. Simmons. Robots in the wild: Observing human-robot social interaction outside the lab. *IEEE International Workshop on Advanced Motion Control*, 2006.

[18] Selma Sabanovic. Robots in society, society in robots. *International Journal of Social Robotics*, 2010.

[19] Vasant Srinivasan and Leila Takayama. Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery, 2016.

[20] Santosh Balajee Banisetty and Tom Williams. Implicit communication through social distancing: Can social navigation communicate social norms? In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[21] Christoforos Mavrogiannis, Alena M. Hutchinson, John Macdonald, Patrícia Alves-Oliveira, and Ross A. Knepper. Effects of distinct robot navigation strategies on human behavior in a crowded environment. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2020.

[22] Yaxin Hu, Yuxiao Qu, Adam Maus, and Bilge Mutlu. Polite or direct? conversation design of a smart display for older adults based on politeness theory. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

[23] Felix Gervits, Gordon Briggs, and Matthias Scheutz. The pragmatic parliament: A framework for socially-appropriate utterance selection in artificial agents. *Cognitive Science*, 2017.

[24] Ruchen Wen, Zhao Han, and Tom Williams. Teacher, teammate, subordinate, friend: Generating norm violation responses grounded in role-based relational norms. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[25] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. Using robots to moderate team conflict: The case of repairing violations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2015.

[26] Anna M. H. Abrams, Pia S. C. Dautzenberg, Carla Jakobowsky, Stefan Ladwig, and Astrid M. Rosenthal-von der Pütten. A theoretical and empirical reflection on technology acceptance models for autonomous delivery robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[27] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. Effects of culture on the credibility of robot speech: A comparison between english and arabic. 2015.

[28] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G Jentsch. Building appropriate trust in human-robot teams. In *AAAI Spring Symposium Series*, 2013.

[29] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. Overtrust of robots in emergency evacuation scenarios. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2016.

[30] Bertram F Malle and Daniel Ullman. A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*. Elsevier, 2021.

[31] A. Jung Moon, Peter Danielson, and H. F.Machiel van der Loos. Survey-Based Discussions on Morally Contentious Applications of Interactive Robotics. *International Journal of Social Robotics*, 2012.

[32] Cathrine Hasse, Stine Trentemøller, and Jessica Sorenson. The Use of Ethnography to Identify and Address Ethical, Legal, and Societal (ELS) Issues. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2018.

[33] Ole Sejer Iversen, Rachel Charlotte Smith, and Christian Dindler. Child as Protagonist: Expanding the Role of Children in Participatory Design. In *Proceedings of the ACM Conference on Interaction Design and Children*, 2017.

[34] Elaheh Sanoubari, John Edison Muñoz Cardona, Hamza Mahdi, James E. Young, Andrew Houston, and Kerstin Dautenhahn. Robots, Bullies and Stories: A Remote Co-design Study with Children. In *Proceedings of the ACM Conference on Interaction Design and Children*, 2021.

[35] Ryan Van Patten, Amber V Keller, Jacqueline E Maye, Dilip V Jeste, Colin Depp, Laurel D Riek, and Elizabeth W Twamley. Home-based cognitively assistive robots: Maximizing cognitive functioning and maintaining independence in older adults without dementia. *Clinical Interventions in Aging*, 2020.

[36] Stephanie Valencia, Michal Luria, Amy Pavel, Jeffrey P Bigham, and Henny Admoni. Co-designing socially assistive sidekicks for motion-based aac. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[37] Julia Galliers, Stephanie Wilson, Abi Roper, Naomi Cocks, Jane Marshall, Sam Muscroft, and Tim Pring. Words are not enough: empowering people with aphasia in the design process. In *Proceedings of the Participatory Design Conference on Research Papers*, 2012.

[38] Amanda Lazar, Jessica L. Feuston, Caroline Edasis, and Anne Marie Piper. Making as Expression: Informing Design with People with Complex Communication Needs through Art Therapy. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2018.

[39] Sanika Moharana, Alejandro E. Panduro, Hee Rin Lee, and Laurel D. Riek. Robots for Joy, Robots for Sorrow: Community Based Robot Design for Dementia Caregivers. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019.

[40] Katie Winkle, Praminda Caleb-Solly, Ailie Turton, and Paul Bremner. Social robots for engagement in rehabilitative therapies: Design implications from a study with therapists. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2018.

[41] Joanna Butchart, Reema Harrison, Jan Ritchie, Felip Martí, Chris McCarthy, Sarah Knight, and Adam Scheinberg. Child and parent perceptions of acceptability and therapeutic value of a socially assistive robot used during pediatric rehabilitation. *Disability and Rehabilitation*, 2021.

[42] Sarah M Rabbitt, Alan E Kazdin, and Brian Scassellati. Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use. *Clinical Psychology Review*, 2015.

[43] Hee Rin Lee, Selma Šabanović, Wan-Ling Chang, Shinichi Nagata, Jennifer Piatt, Casey Bennett, and David Hakken. Steps toward participatory design of social robots: Mutual learning with older adults with depression. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[44] Caroline L. van Straten, Jochen Peter, Rinaldo Kahne, and Alex Barco. The wizard and i: How transparent teleoperation and self-description (do not) affect children's robot perceptions and child-robot relationship formation. *AI & Society*, 2021.

[45] Laetitia Tanqueray, Tobiaz Paulsson, Mengyu Zhong, Stefan Larsson, and Ginevra Castellano. Gender fairness in social robotics: Exploring a future care of peripartum depression. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[46] Elin Björling and Emma Rose. Participatory Research Principles in Human-Centered Design: Engaging Teens in the Co-Design of a Social Robot. *Multimodal Technologies and Interaction*, 2019.

[47] Khon Anastasia, Raechel Walker, Madhurima Das, Maria Yang, Cynthia Breazea, Hae Won Park, and Aditi Verma. Ethics, equity, & justice in human-robot interaction: A review and future directions. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2022.

[48] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021.

[49] Solace Shen, Petr Slovak, and Malte F. Jung. "stop. i see a conflict happening.": A robot mediator for young children's interpersonal conflict resolution. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2018.

[50] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. Using robots to moderate team conflict: The case of repairing violations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2015.

[51] Takanori Komatsu, Bertram F. Malle, and Matthias Scheutz. Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across u.s. and japan. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[52] Victoria Groom, Jimmy Chen, Theresa Johnson, F. Arda Kara, and Clifford Nass. Critic, compatriot, or chump? responses to robot blame attribution. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010.

[53] Alyssa Hanson, Nichole Starr, Cloe Emnett, Ruchen Wen, Bertram Malle, and Tom Williams. The power of advice: Differential blame for human and robot advisors and deciders in a moral advising context human and robot advisors and deciders in a moral advising context. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2024.

[54] Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *Transactions on Interactive Intelligent Systems*, 2018.

[55] Hideki Garcia, Katie Winkle, Tom Williams, and Megan Strait. Victims and observers: How gender, victimization experience, and biases shape perceptions of robot abuse. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2023.

[56] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Bršcić, Iolanda Leite, and Tom Williams. Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[57] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. Tact in noncompliance: The need for pragmatically apt responses to unethical commands. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

[58] Terran Mott and Tom Williams. Confrontation and cultivation: Understanding perspectives on robot responses to norm violations. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2023.

[59] Benedict Tay, Younbo Jung, and Taezoon Park. When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior*, 2014.

[60] Clifford Nass, Youngme Moon, and Nancy Green. Are machines gender neutral? gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, 1997.

[61] Ryan Blake Jackson and Tom Williams. Language-capable robots may inadvertently weaken human moral norms. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019.

[62] Tom Williams, Ryan Jackson, and Jane Lockshin. A bayesian analysis of moral norm malleability during clarification dialogues. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2018.

[63] Qin Zhu, Tom Williams, and Ruchen Wen. Role-based morality, ethical pluralism, and morally capable robots. *Journal of Contemporary Eastern Asia*, 2021.

[64] Ruchen Wen and Tom Williams. Hidden complexities in the computational modeling of proportionality for robotic norm violation response. In *AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (AI-HRI)*, 2022.

[65] Jenay M Beer, Arthur D Fisk, and Wendy A Rogers. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction*, 2014.

[66] Mica R Endsley and David B Kaber. Level of automation effects on performance, situation awareness, and workload in a dynamic control task. *Ergonomics*, 1999.

[67] Clifford D Johnson, Michael E Miller, Christina F Rusnock, and David R Jacques. A framework for understanding automation in terms of levels of human control abstraction. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017.

[68] Amro Khasawneh, Hunter Rogers, Jeffery Bertrand, Kapil Chalil Madathil, and Anand Gramopadhye. Human adaptation to latency in teleoperated multi-robot human-agent search and rescue teams. *Automation in Construction*, 2019.

[69] Lu Feng, Clemens Wiltsche, Laura Humphrey, and Ufuk Topcu. Synthesis of human-in-the-loop control protocols for autonomous systems. *Transactions on Automation Science and Engineering*, 2016.

[70] Taemie Kim and Pamela Hinds. Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2006.

[71] Mica R Endsley. Situation awareness in future autonomous vehicles: Beware of the unexpected. In *Congress of the International Ergonomics Association*. Springer, 2018.

[72] Jessie Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. Situation awareness–based agent transparency. Technical report, US Army Research Laboratory, 2014.

[73] Brian Scassellati, Henny Admoni, and Maja Matarić. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 2012.

[74] Saad Elbeleidy, Terran Mott, and Tom Williams. Practical, ethical, and overlooked: Teleoperated socially assistive robots in the quest for autonomy. In *Companion Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*, 2022.

[75] Peter H. Kahn and Solace Shen. *NOC NOC, Who's There? A New Ontological Category (NOC) for Social Robots.* Cambridge University Press, 2017.

[76] Ryan Blake Jackson and Tom Williams. A theory of social agency for human-robot interaction. *Frontiers in Robotics and AI*, 2021.

[77] Herbert Clark and Kerstin Fischer. Social robots as depictions of social agents. *Behavioral and Brain Sciences*, 2022.

[78] Kara Weisman. Extraordinary entities: Insights into folk ontology from studies of lay people's beliefs about robots. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2022.

[79] Eleonore Lumer and Hendrik Buschmeier. Should robots be polite? expectations about politeness in human–robot interaction. *Frontiers in Robotics and AI*, 2023.

[80] Eleonore Lumer and Hendrik Buschmeier. Perception of power and distance in human-human and human-robot role-based relations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[81] Penelope Brown and Stephen C Levinson. *Politeness: Some Universals in Language Usage.* Cambridge University Press, 1987.

[82] Imran Fanaswala, Brett Browning, and Majd Sakr. Interactional disparities in english and arabic native speakers with a bi-lingual robot receptionist. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2011.

[83] Diede P.M. Van der Hoorn, Anouk Neerincx, and Maartje M.A. de Graaf. I think you are doing a bad job!": The effect of blame attribution by a robot in human-robot collaboration. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[84] Goren Gordon. Social behaviour as an emergent property of embodied curiosity: a robotics perspective. *Philosophical Transactions of the Royal Society of London*, 2019.

[85] Thomas Holtgraves. Understanding miscommunication: Speech act recognition in digital contexts. *Cognitive Science*, 2021.

[86] Leigh Clark, Abdulmalik Yusuf Ofemile, and Benjamin Cowan. *Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech.* Springer, 2020.

[87] Autumn Edwards, Chad Edwards, and Andrew Gambino. The social pragmatics of communication with social robots: Effects of robot message design logic in a regulative context. *International Journal of Social Robotics*, 2020.

[88] Tom Williams, Priscilla Briggs, and Matthias Scheutz. Covert robot-robot communication: Human perceptions and implications for human-robot interaction. *Journal of Human-Robot Interaction*, 2015.

[89] Leigh Clark. Social boundaries of appropriate speech in hci: A politeness perspective. In *Proceedings of British HCI*, 2018.

[90] David Feil-Seifer and Maja J Mataric. Defining socially assistive robotics. In *Proceedings of the IEEE International Conference on Rehabilitation Robotics*, 2005.

[91] Juan Fasola and Maja J Mataric. Using socially assistive human–robot interaction to motivate physical exercise for older adults. *Proceedings of the IEEE*, 2012.

[92] Kazuyoshi Wada, Takanori Shibata, and Yukitaka Kawaguchi. Long-term robot therapy in a health service facility for the aged: A case study for 5 years. *International Conference on Rehabilitation Robotics*, 2009.

[93] Joshua J Diehl, Lauren M Schmitt, Michael Villano, and Charles R Crowell. The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in Autism Spectrum Disorders*, 2012.

[94] Mohammed A Saleh, Habibah Hashim, Nur Nabila Mohamed, Ali Abd Almisreb, and Benjamin Durakovic. Robots and autistic children: A review. *Periodicals of Engineering and Natural Sciences (PEN)*, 2020.

[95] Zohreh Salimi, Ensiyeh Jenabi, and Saeid Bashirian. Are social robots ready yet to be used in care and therapy of autism spectrum disorder: A systematic review of randomized controlled trials. *Neuroscience & Biobehavioral Reviews*, 2021.

[96] John-John Cabibihan, Hifza Javed, Marcelo Ang, and Sharifah Mariam 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*, 2013.

[97] Fabiane Barreto Vavassori Benitti. Exploring the educational potential of robotics in schools: A systematic review. *Computers & Education*, 2012.

[98] David Silvera-Tawil and Christine Roberts-Yates. Socially-assistive robots to enhance learning for secondary students with intellectual sisabilities and autism. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018.

[99] Layne Jackson Hubbard, Yifan Chen, Eliana Colunga, Pilyoung Kim, and Tom Yeh. Child-robot interaction to integrate reflective storytelling into creative play. In *Proceedings of the ACM Conference on Creativity and Cognition*, 2021.

[100] Madeline M Blankenship and Cathy Bodine. Socially assistive robots for children with cerebral palsy: A meta-analysis. *IEEE Transactions on Medical Robotics and Bionics*, 2020.

[101] Katarzyna Kabacińska, Tony J Prescott, and Julie M Robillard. Socially assistive robots as mental health interventions for children: A scoping review. *International Journal of Social Robotics*, 2021.

[102] Katta Spiel, Christopher Frauenberger, and Geraldine Fitzpatrick. Experiences of autistic children with technologies. *International Journal of Child-Computer Interaction*, 2017.

[103] Jason C. Yip, Kiley Sobel, Caroline Pitt, Kung Jin Lee, Sijin Chen, Kari Nasu, and Laura R. Pina. Examining adult-child interactions in intergenerational participatory design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2017.

[104] Ole Sejer Iversen and Christina Brodersen. Building a bridge between children and users: A socio-cultural approach to child-computer interaction. *Cognition, Technology, and Work*, 2008.

[105] Terran Mott, Alexandra Bejarano, and Tom Williams. Robot co-design can help us engage child stakeholders in ethical reflection. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[106] Janet C. Read, Matthew Horton, Gavin Sim, Peggy Gregory, Daniel Fitton, and Brendan Cassidy. Check: A tool to inform and encourage ethical practice in participatory design with children. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 2013.

[107] Maarten Van Mechelen, Line Have Musaeus, Ole Sejer Iversen, Christian Dindler, and Arthur Hjorth. A systematic review of empowerment in child-computer interaction research. In *Proceedings of the ACM Interaction Design and Children Conference*, 2021.

[108] Jacqueline M. Kory-Westlund and Cynthia Breazeal. Exploring the effects of a social robot's speech entrainment and backstory on young children's emotion, rapport, relationship, and learning. *Frontiers in Robotics and AI*, 2019.

[109] Jacqueline M. Kory Westlund, Hae Won Park, Randi Williams, and Cynthia Breazeal. Measuring young children's long-term relationships with social robots. In *Proceedings of the ACM Conference on Interaction Design and Children*, 2018.

[110] Rebecca Stower, Natalia Calvo Barajas, Ginevra Castellano, and Arvid Kappas. A meta-analysis on children's trust in social robots. *International Journal of Social Robotics*, 2021.

[111] Gail F. Melson, Peter H. Kahn, Alan M. Beck, Batya Friedman, Trace Roberts, and Erik Garrett. Robots as dogs? children's interactions with the robotic dog aibo and a live australian shepherd. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 2005.

[112] Jacqueline M. Kory-Westlund and Cynthia Breazeal. Deception, secrets, children, and robots: What's acceptable? *Proceedings of the workshop on The Emerging Policy and Ethics of Human-Robot Interaction at HRI2015*, 2015.

[113] Caitlyn Clabaugh and Maja Matarić. Escaping oz: Autonomy in socially assistive robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 2019.

[114] François Michaud, Tamie Salter, Audrey Duquette, Henri Mercier, Michel Lauria, Helene Larouche, and Francois Larose. Assistive technologies and child-robot interaction. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, 2007.

[115] PEERbots. Peerbots, 2021. URL https://peerbots.org.

[116] Fine Art Miracles Inc. Fine art miracles, 2021. URL https://fineartmiracles.com/.

[117] Misty Robotics. Misty robotics, 2022. URL https://www.mistyrobotics.com/.

[118] Movia Robotics. Movia robotics, 2021. URL https://moviarobotics.com/.

[119] Maja J Matarić and Brian Scassellati. Socially assistive robotics. In *Springer Handbook of Robotics*. Springer, 2016.

[120] Laura Boccanfuso, Sarah Scarborough, Ruth K Abramson, Alicia V Hall, Harry H Wright, and Jason 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*, 2017.

[121] Munjal Desai, Kristen Stubbs, Aaron Steinfeld, and Holly Yanco. Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of AISB '09 Convention: New Frontiers in Human-Robot Interaction*, 2009.

[122] Elaine C Lu, Rosalie H Wang, Debbie Hebert, Jennifer Boger, Mary P Galea, and Alex Mihailidis. The development of an upper limb stroke rehabilitation robot: Identification of clinical practices and design requirements through a survey of therapists. *Disability and Rehabilitation: Assistive Technology*, 2011.

[123] Roman Kulikovskiy, Megan Sochanski, Matteson Eaton, Jessica Korneder, Wing-Yue Geoffrey Louie, et al. Can therapists design robot-mediated interventions and teleoperate robots using vr to deliver interventions for asd? In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[124] Stanislav Ivanov and Craig Webster. What should robots do? a comparative analysis of industry professionals, educators, and tourists. In *Information and Communication Technologies in Tourism*. 2019.

[125] Saad Elbeleidy, Terran Mott, Dan Liu, Ellen Yi-Luen Do, Elizabeth Reddy, and Tom Williams. Beyond the session: Centering teleoperators in robot-assisted therapy reveals the bigger picture. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2023.

[126] Saad Elbeleidy, Terran Mott, Dan Liu, and Tom Williams. Practical considerations for deploying robot teleoperation in therapy and telehealth. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2022.

[127] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. Social robots for education: A review. *Science robotics*, 2018.

[128] Laurie A Dickstein-Fischer, Darlene E Crone-Todd, Ian M Chapman, Ayesha T Fathima, and Gregory S Fischer. Socially assistive robots: Current status and future prospects for autism interventions. *Innovation and Entrepreneurship in Health*, 2018.

[129] Ester Martinez-Martin, Felix Escalona, and Miguel Cazorla. Socially assistive robots for older adults and people with autism: An overview. *Electronics*, 2020.

[130] Junya Nakanishi, Jun Baba, and Hiroshi Ishiguro. Robot-mediated interaction between children and older adults: A pilot study for greeting tasks in nursery schools. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2022.

[131] Katarzyna Kabacińska, Tony J Prescott, and Julie M Robillard. Socially assistive robots as mental health interventions for children: A scoping review. *International Journal of Social Robotics*, 2020.

[132] Kenneth Barish. What is therapeutic in child therapy? i. therapeutic engagement. *Psychoanalytic Psychology*, 2004.

[133] Martin Cooney and Maria Menezes. Design for an art therapy robot: An explorative review of the theoretical foundations for engaging in emotional and creative painting with a robot. *Multimodal Technologies and Interaction*, 2018.

[134] Hideki Takahashi, Kanae Matsushima, and Toshihiro Kato. The effectiveness of dance/movement therapy interventions for autism spectrum disorder: A systematic review. *American Journal of Dance Therapy*, 2019.

[135] Thomas Armstrong. *Neurodiversity: Discovering the Extraordinary Gifts of Autism, ADHD, Dyslexia, and Other Brain Differences.* Hachette Books, 2010.

[136] Luthffi Idzhar Ismail, Thibault Verhoeven, Joni Dambre, and Francis Wyffels. Leveraging robotics research for children with autism: A review. *International Journal of Social Robotics*, 2019.

[137] Micki McGee. Neurodiversity. *Contexts*, 2012.

[138] Kristen Bottema-Beutel, Steven K Kapp, Jessica Nina Lester, Noah J Sasson, and Brittany N Hand. Avoiding ableist language: Suggestions for autism researchers. *Autism in Adulthood*, 2021.

[139] Thomas Armstrong. Neurodiversity: The future of special education? *Educational Leadership*, 2017.

[140] Autistic Self Advocacy Network and L. Berry. *Welcome to the Autistic Community.* Autistic Press, 2020.

[141] Robert Chapman. Neurodiversity theory and its discontents: Autism, schizophrenia, and the social model of disability. *The Bloomsbury Companion to Philosophy of Psychiatry*, 2019.

[142] Guy Dewsbury, Karen Clarke, Dave Randall, Mark Rouncefield, and Ian Sommerville. The anti-social model of disability. *Disability & Society*, 2004.

[143] M Blamires. Towards an educational model for pupils with autism and asperger's syndrome. In *Children with Learning Difficulties: A Collaborative Approach to their Education and Management.* Wiley, 1997.

[144] Joshua Wainer, Ester Ferrari, Kerstin Dautenhahn, and Ben Robins. The effectiveness of using a robotics class to foster collaboration among groups of children with autism in an exploratory study. *Personal and Ubiquitous Computing*, 2010.

[145] Irini Giannopulu. Multimodal cognitive nonverbal and verbal interactions: The neurorehabilitation of autistic children via mobile toy robots. *IARIA International Journal of Advances in Life Sciences*, 2013.

[146] Elizabeth S Kim, Lauren D Berkovits, Emily P Bernier, Dan Leyzberg, Frederick Shic, Rhea Paul, and Brian Scassellati. Social robots as embedded reinforcers of social behavior in children with autism. *Journal of Autism and Developmental Disorders*, 2013.

[147] Eva Yin-han Chung. Robotic intervention program for enhancement of social engagement among children with autism spectrum disorder. *Journal of Developmental and Physical Disabilities*, 2019.

[148] Oliver Damm, Karoline Malchus, Petra Jaecks, Soeren Krach, Frieder Paulus, Marnix Naber, Andreas Jansen, Inge Kamp-Becker, Wolfgang Einhaeuser-Treyer, Prisca Stenneken, et al. Different gaze behavior in human-robot interaction in asperger's syndrome: An eye-tracking study. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2013.

[149] Flannery Hope Currin, Kyle Diederich, Kaitlyn Blasi, Allyson Dale Schmidt, Holly David, Kerry Peterman, and Juan Pablo Hourcade. Supporting shy preschool children in joining social play. In *Proceedings of the ACM Conference on Interaction Design and Children*, 2021.

[150] Joshua Wainer, Ben Robins, Farshid Amirabdollahian, and Kerstin Dautenhahn. Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism. *IEEE Transactions on Autonomous Mental Development*, 2014.

[151] Eva Yin-han Chung. Robot-mediated social skill intervention programme for children with autism spectrum disorder: An aba time-series study. *International Journal of Social Robotics*, 2020.

[152] Diego Zapata-Rivera, Tanner Jackson, and I Katz. Authoring conversation-based assessment scenarios. *Design Recommendations for Intelligent Tutoring Systems*, 2015.

[153] Michael Villano, Charles R Crowell, Kristin Wier, Karen Tang, Brynn Thomas, Nicole Shea, Lauren M Schmitt, and Joshua 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 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011.

[154] Patrick Gebhard, Gregor Mehlmann, and Michael Kipp. Visual scenemaker—a tool for authoring interactive virtual characters. *Journal on Multimodal User Interfaces*, 2012.

[155] Linbo Luo, Wentong Cai, Suiping Zhou, Michael Lees, and Haiyan Yin. A review of interactive narrative systems and technologies: A training perspective. *Simulation*, 2015.

[156] Jennifer Carlson, Robin R Murphy, and Andrew Nelson. Follow-up analysis of mobile robot failures. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2004.

[157] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 2000.

[158] Raja Parasuraman and Christopher D Wickens. Humans: Still vital after all these years of automation. *Human Factors*, 2008.

[159] Priyesh Tiwari, Jim Warren, Karen J Day, and Bruce MacDonald. Some non-technology implications for wider application of robots to assist older people. *Health Care and Informatics Review Online*, 2010.

[160] Sebastian Thrun. Toward a framework for human-robot interaction. *Human Computer Interaction*, 2004.

[161] Maja J Matarić. Socially assistive robotics: Human augmentation versus automation. *Science Robotics*, 2017.

[162] Ylona Chun Tie, Melanie Birks, and Karen Francis. Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, 2019.

[163] Paul Atkinson, Amanda Coffey, and Sara Delamont. *Key Themes in Qualitative Research: Continuities and Changes.* Rowman Altamira, 2003.

[164] Kathy Charmaz. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis.* SAGE Publications, 2006.

[165] Laura L Downey. Group usability testing: Evolution in usability techniques. *Journal of Usability Studies*, 2007.

[166] Teal W Benevides, Stephen M Shore, Kate Palmer, Patricia Duncan, Alex Plank, May-Lynn Andresen, Reid Caplan, Barb Cook, Dena Gassner, Becca Lory Hector, et al. Listening to the autistic voice: Mental health priorities to guide research and practice in autism from a stakeholder-driven project. *Autism*, 2020.

[167] Richard L Simpson, Michael McKee, Dixie Teeter, and Alyson Beytien. Evidence-based methods for children and youth with autism spectrum disorders: Stakeholder issues and perspectives. *Exceptionality*, 2007.

[168] Richard Chasin and Tanya B White. The child in family therapy: Guidelines for active engagement across the age span. *Children in Family Contexts: Perspectives on Treatment*, 1989.

[169] Maureen Werrbach. Common compensation models for group practices, 2019. URL blog.therapynotes.com/common-compensation-models-for-group-practices.

[170] Dagoberto Cruz-Sandoval, Arturo Morales-Tellez, Eduardo Benitez Sandoval, and Jesus 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*, 2020.

[171] Kirsti Malterud, Volkert Dirk Siersma, and Ann Dorrit Guassora. Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, 2016.

[172] Robert B Cialdini and Melanie R Trost. Social influence: Social norms, conformity, and compliance. In *The Handbook of Social Psychology*. McGraw-Hill, 1998.

[173] Ryan Blake Jackson, Tom Williams, and Nicole Smith. Exploring the role of gender in perceptions of robotic noncompliance. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020.

[174] Katie Seaborn and Peter Pennefather. Gender neutrality in robots: An open living review framework. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[175] Ruchen Wen, Mohammed Aun Siddiqui, and T. Williams. Dempster-shafer theoretic learning of indirect speech act comprehension norms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[176] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. The impact of different ethical frameworks underlying a robot's advice on charitable donations. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2023.

[177] Danette Ifert Johnson, Michael E. Roloff, and Melissa A. Riffee. Politeness theory and refusals of requests: Face threat as a function of expressed obstacles. *Communication Studies*, 2004.

[178] Erving Goffman. *Interaction Ritual: Essays in Face-to-Face Behavior*. Routledge, 1967.

[179] Marilyn A. Walker, Janet E. Cahn, and Stephen J. Whittaker. Improvising linguistic style: Social and affective bases for agent personality. In *AAMAS*, 1997.

[180] Sahba Zojaji, Christopher Peters, and Catherine Pelachaud. Influence of virtual agent politeness behaviors on how users join small conversational groups. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, 2020.

[181] Swati Gupta, Marilyn A. Walker, and Daniela M. Romano. Generating politeness in task based interaction: An evaluation of the effect of linguistic form and culture. In *Proceedings of the European Workshop on Natural Language Generation*, 2007.

[182] Maha Salem, Micheline Ziadee, and Majd Sakr. Marhaba, how may i help you? effects of politeness and culture on robot acceptance and anthropomorphization. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2014.

[183] Sara Mills. Gender and impoliteness. *Journal of Politeness Research*, 2005.

[184] Cailyn Smith, Charlotte Gorgemans, Ruchen Wen, Saad Elbeleidy, Sayanti Roy, and Tom Williams. Leveraging intentional factors and task context to predict linguistic norm adherence. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2022.

[185] Erin R. Hoffman, David W. McDonald, and Mark Zachry. Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. *Proceedings of the ACM on Human-Computer Interaction*, 2017.

[186] Nasif Imtiaz, Justin Middleton, Peter Girouard, and Emerson Murphy-Hill. Sentiment and politeness analysis tools on developer discussions are unreliable, but so are people. In *Proceedings of the International Workshop on Emotion Awareness in Software Engineering*, 2018.

[187] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. Intimate futures: Staying with the trouble of digital personal assistants through design fiction. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)*, 2018.

[188] Richmond Y. Wong. Using design fiction memos to analyze ux professionals' values work practices: A case study bridging ethnographic and design futuring methods. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.

[189] Qin Zhu, Tom Williams, and Ryan Jackson. Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective. Springer, 2018.

[190] Gordon Briggs, Tom Williams, Ryan Blake Jackson, and Matthias Scheutz. Why and how robots should say 'no'. *International Journal of Social Robotics*, 2022.

[191] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[192] Ryan Blake Jackson, Sihui Li, Santosh Balajee Banisetty, Sriram Siva, Hao Zhang, Neil Dantam, and Tom Williams. An integrated approach to context-sensitive moral cognition in robot cognitive architectures. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[193] Gordon Briggs and Matthias Scheutz. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal Social Robotics*, 2014.

[194] Sharan B Merriam et al. Introduction to qualitative research. *Qualitative Research in Practice*, 2002.

[195] Clive Seale. Quality in Qualitative Research. *Qualitative Inquiry*, 1999.

[196] Hee Rin Lee, EunJeong Cheon, Chaeyun Lim, and Kerstin Fischer. Configuring humans: What roles humans play in hri research. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[197] Alan Borning and Michael Muller. Next steps for value sensitive design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2012.

[198] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, 2020.

[199] Eric P.S. Baumer, Timothy Berrill, Sarah C. Botwinick, Jonathan L. Gonzales, Kevin Ho, Allison Kundrik, Luke Kwon, Tim LaRowe, Chanh P. Nguyen, Fredy Ramirez, Peter Schaedler, William Ulrich, Amber Wallace, Yuchen Wan, and Benjamin Weinfeld. What would you do? design fiction and ethics. In *Proceedings of the ACM International Conference on Supporting Group Work (GROUP*, 2018.

[200] Richmond Y. Wong, Ellen Van Wyk, and James Pierce. Real-fictional entanglements: Using science fiction and design fiction to interrogate sensing technologies. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)*, 2017.

[201] Michal Luria and Stuart Candy. Letters from the future: Exploring ethical dilemmas in the design of social agents. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

[202] Sarah Fox, Noura Howell, Richmond Wong, and Franchesca Spektor. Vivewell: Speculating near-future menstrual tracking through current data practices. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)*, 2019.

[203] Janet Shibley Hyde, Rebecca S Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M van Anders. ”the future of sex and gender in psychology: Five challenges to the gender binary. *The American Psychologist*, 2019.

[204] Mark West, Rebecca Kraut, and Chew Han Ei. I’d blush if i could: Closing gender divides in digital skills through education. UNESCO, 2019.

[205] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. The shape of our bias: Perceived age and gender in the humanoid robots of the abot database. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[206] Icek Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 1991.

[207] Mark Coeckelbergh. Robot rights? towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 2010.

[208] Sven Nyholm. *Humans and Robots: Ethics, Agency, and Anthropomorphism.* Rowman & Littlefield Publishers, 2020.

[209] Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 2004.

[210] Anna M. H. Abrams, Pia S. C. Dautzenberg, Carla Jakobowsky, Stefan Ladwig, and Astrid M. Rosenthal-von der Pütten. A Theoretical and Empirical Reflection on Technology Acceptance Models for Autonomous Delivery Robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[211] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, 2015.

[212] Vanessa Evers, Heidy Maldonado, Talia Brodecki, and Pamela Hinds. Relational vs. Group Self-Construal: Untangling the Role of National Culture in HRI. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2008.

[213] Ryan Blake Jackson and Tom Williams. Enabling Morally Sensitive Robotic Clarification Requests. *ACM Transactions on Human-Robot Interaction*, 2022.

[214] Danette Ifert Johnson, Michael Roloff, and Melissa Riffee. Responses to refusals of requests: Face threat and persistence, persuasion, and forgiving statements. *Communication Quarterly*, 2004.

[215] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Daniel Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. In *Annual Meeting of the Association for Computational Linguistics*, 2013.

[216] Danette Ifert Johnson. Politeness theory and conversational refusals: Associations between various types of face threat and perceived competence. *Western Journal of Communication*, 2007.

[217] Richard J. Watts. *Politeness.* Cambridge University Press, 2003.

[218] Marina Terkourafi. Beyond the micro-level in politeness research. *Journal of Politeness Research-Language Behaviour Culture*, 2005.

[219] Geoffrey Leech. *The Pragmatics of Politeness.* Oxford University Press, 2014.

[220] Stephan Hammer, Birgit Lugrin, Sergey Bogomolov, Kathrin Janowski, and Elisabeth André. Investigating politeness strategies and their persuasiveness for a robotic elderly assistant. In *Proceedings of the 11th International Conference on Persuasive Technology*. Springer-Verlag, 2016.

[221] Namyeon Lee, Jeonghun Kim, Eunji Kim, and Ohbyung Kwon. The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *International Journal of Social Robotics*, 2017.

[222] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2019.

[223] Caroline Pantofaru, Leila Takayama, Tully Foote, and Bianca Soto. Exploring the role of robots in home organization. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012.

[224] Christoforos Mavrogiannis, Alena M. Hutchinson, John Macdonald, Patricia Alves-Oliveira, and Ross A. Knepper. Effects of Distinct Robot Navigation Strategies on Human Behavior in a Crowded Environment. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019.

[225] De'Aira Bryant, Jason Borenstein, and Ayanna Howard. Why Should We Gender?: The Effect of Robot Gendering and Occupational Stereotypes on Human Trust and Perceived Competency. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2020.

[226] Steven Wilson and Adrianne Kunkel. Identity implications of influence goals: Similarities in perceived face threats and facework across sex and close relationships. *Journal of Language and Social Psychology*, 2000.

[227] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 2018.

[228] Terran Mott, Aaron Fanganello, and Tom Williams. What a thing to say! which linguistic politeness strategies should robots use in noncompliance interactions? In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2024.

[229] Don van den Bergh, Johnny Van Doorn, Maarten Marsman, Tim Draws, Erik-Jan Van Kesteren, Koen Derks, Fabian Dablander, Quentin F Gronau, Šimon Kucharskỳ, Akash R Komarlu Narendra Gupta, et al. A tutorial on conducting and interpreting a bayesian anova in jasp. *L'Année psychologique*, 2020.

[230] JASP Team. JASP (Version 0.18.0)[Computer software], 2023. URL https://jasp-stats.org/.

[231] Dominique Makowski, Mattan Ben-Shachar, and Daniel Lüdecke. bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 2019.

[232] Richard Morey, Jeffrey Rouder, Tahira Jamil, and Maintainer Richard Morey. Package 'bayesfactor'. *URL http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor*, 2015.

[233] Michael D Lee and Eric-Jan Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2014.

[234] Harold Jeffreys. *Theory of Probability*. The International Series of Monographs on Physics. Clarendon Press, 1948.

[235] Joseph Simmons, Leif Nelson, and Uri Simonsohn. False-positive psychology. *Psychological Science*, 2011.

[236] Eric-Jan Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 2007.

[237] Jonathan Sterne and George Davey Smith. Sifting the evidence - what's wrong with significance tests? *British Medical Journal (Clinical Research Edition))*, 2001.

[238] Andrew Jarosz and Jennifer Wiley. What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, 2014.

[239] A J Verhagen and Eric-Jan Wagenmakers. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology. General*, 2014.

[240] Alexander Ly, Alexander Etz, Maarten Marsman, and Eric-Jan Wagenmakers. Replication bayes factors from evidence updating. *Behavior Research Methods*, 2018.

[241] Aidan Naughton and Tom Williams. How to tune your draggin': Can body language mitigate face threat in robotic noncompliance? In *Proceedings of the International Conference on Social Robotics (ICSR)*. Springer, 2021.

[242] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *Autonomous Agents and Multi-Agent Systems (AAMAS)*. Springer, 2019.

[243] Victoria Alonso and Paloma de la Puente. System transparency in shared autonomy: A mini review. *Frontiers in Neurorobotics*, 2018.

[244] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. Explainable embodied agents through social cues: A review. *Transactions on Human Robot Interaction*, 2021.

[245] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics Guidelines for Trustworthy AI.* EU Publications Office, 2019.

[246] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. What does the robot think? transparency as a fundamental design requirement for intelligent systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2016.

[247] Serena Booth, Sanjana Sharma, Sarah Chung, Julie Shah, and Elena L. Glassman. Revisiting human-robot teaching and learning through the lens of human concept learning. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[248] Minae Kwon, Malte F. Jung, and Ross A. Knepper. Human expectations of social robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

[249] Avi Rosenfeld and Ariella Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 2019.

[250] Tom Williams. The eye of the robot beholder: Ethical risks of representation, recognition, and reasoning over identity characteristics in human-robot interaction. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*, 2023.

[251] Elin A. Björling, Emma Rose, and Rachel Ren. Teen-Robot Interaction: A Pilot Study of Engagement with a Low-fidelity Prototype. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, 2018.

[252] Elin A. Björling and Emma Rose. Participatory Research Principles in Human-Centered Design: Engaging Teens in the Co-Design of a Social Robot. *Multimodal Technologies and Interaction*, 2019.

[253] Tom Williams. Understanding roboticists' power through matrix guided technology power analysis. In *Companion Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*, 2024.

[254] Terran Mott and Tom Williams. How can dog handlers help us understand the future of wilderness search & rescue robots? In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2023.

[255] Giulia Perugia and Dominika Lisy. Robot's gendering trouble: A scoping review of gendering humanoid robots and its effects on hri. *International Journal of Social Robotics*, 2022.

[256] Katie Winkle, Donald McMillan, Maria Arnelid, Katherine Harrison, Madeline Balaam, Ericka Johnson, and Iolanda Leite. Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023.

[257] Chloe McCaffrey, Alexander Taylor, Sayanti Roy, Santosh Balajee Banisetty, Ross Mead, and Tom Williams. Can Robots Be Used to Encourage Social Distancing? In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 2021.

[258] Guy Hoffman, Oren Zuckerman, Gilad Hirschberger, Michal Luria, and Tal Shani Sherman. Design and evaluation of a peripheral robotic conversation companion. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015.

[259] Hadas Erel, Yoav Cohen, Klil Shafrir, Sara Daniela Levy, Idan Dov Vidra, Tzachi Shem Tov, and Oren Zuckerman. Excluded by robots: Can robot-robot-human interaction lead to ostracism? In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[260] Duri Long, Takeria Blunt, and Brian Magerko. Co-Designing AI Literacy Exhibits for Informal Learning Spaces. *Proceedings of the ACM on Human-Computer Interaction*, 2021.

[261] David Porfirio, Evan Fisher, Allison Sauppé, Aws Albarghouthi, and Bilge Mutlu. Bodystorming human-robot interactions. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, 2019.

[262] Katie Winkle, Emmanuel Senft, and Séverin Lemaignan. Leador: A method for end-to-end participatory design of autonomous social robots, 2021.

[263] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior acm reference format. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[264] Ruchen Wen, Boyoung Kim, Elizabeth Phillips, Qin Zhu, and Tom Williams. On further reflection... moral reflections enhance robotic moral persuasive capability. In *Proceedings of the International Conference on Persuasive Technology*, 2023.

[265] Boyoung Kim, Ruchen Wen, Ewart Visser, Chad Tossell, Qin Zhu, Tom Williams, and Elizabeth Phillips. Can robot advisers encourage honesty?: Considering the impact of rule, identity, and role-based moral advice. *International Journal of Human-Computer Studies*, 2024.

[266] Ruchen Wen, Boyoung Kim, Elizabeth Phillips, Qin Zhu, and Tom Williams. Comparing norm-based and role-based strategies for robot communication of role-grounded moral norms. *Transactions on Human Robot Interaction*, 2023.

[267] Simson Garfinkel, Jeanna Matthews, Stuart S. Shapiro, and Jonathan M. Smith. Toward algorithmic transparency and accountability. *Communications of the Association for Computing Machinery*, 2017.

[268] Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò Larrieux. Robots and transparency: The multiple dimensions of transparency in the context of robot technologies. *IEEE Robotics & Automation Magazine*, 2019.

[269] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. Dark patterns of explainability, transparency, and user control for intelligent systems. In *Workshops at the ACM Conference on Intelligent User Interfaces*, 2019.

[270] HE Gaole, Web, and Ujwal Gadiraju. Walking on eggshells: Using analogies to promote appropriate reliance in human-ai decision making to in human-ai. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

[271] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 2019.

[272] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[273] Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Anita Sun, Alison Hu, Kate Glazko, Nikos Arechiga, Matthew Lee, Scott Carter, and Haiyi Zhuand Kenneth Holstein. Towards a learner-centered explainable ai. 2022.

[274] Scott Ososky, Tracy Sanders, Florian Jentsch, Peter Hancock, and Jessie Chen. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. 2014.

[275] Maartje de Graaf and Bertram Malle. How people explain action (and ais should too). In *Artificial Intelligence for Human-Robot Interaction AAAI Technical Report*. AAAI, 2017.

[276] Johannes Kraus, Franziska Babel, Philipp Hock, Katrin Hauber, and Martin Baumann. The trustworthy and acceptable hri checklist (ta-hri): Questions and design recommendations to support a trustworthy and acceptable design of human-robot interaction. *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, 2022.

[277] Gail F. Melson, Peter H. Kahn, Alan M. Beck, Batya Friedman, Trace Roberts, and Erik Garrett. Robots as dogs? children's interactions with the robotic dog aibo and a live australian shepherd. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 2005.

[278] Elena Kokkoni, Amanda J. Arnold, Kleio Baxevani, and Herbert G. Tanner. Infants respond to robot's need for assistance in pursuing action-based goals. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2021.

[279] Kerstin Fischer. When transparent does not mean explainable. In *Proceedings of the workshop on Explainable Robotic Systems at HRI2017*, 2017.

[280] Maria Luce Lupetti. Designing playful hri: Acceptability of robots in everyday life through play. In *Proceedings of the ACM/IEEE International Conference on Human Robot Interaction (HRI)*, 2016.

[281] William Gaver. Designing for homo ludens. *I3 Magazine*, 2002.

[282] Wen-Ying Lee and Malte Jung. Ludic-hri: Designing playful experiences with robots. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2020.

[283] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES*, 2021.

[284] Ryan Blake Jackson, Sihui Li, Santosh Balajee Banisetty, Sriram Siva, Hao Zhang, Neil Dantam, and Tom Williams. An integrated approach to context-sensitive moral cognition in robot cognitive architectures. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[285] Matthias Scheutz, T. Williams, Evan A. Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler M. Frasca. An overview of the distributed integrated cognition affect and reflection diarc architecture. *Intelligent Systems, Control and Automation: Science and Engineering*, 2018.

[286] Ehud Reiter. Has a consensus natural language generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*. Association for Computational Linguistics, 1994.

[287] Tom Williams, Cynthia Matuszek, Ross Mead, and Nick Depalma. Scarecrows in oz: The use of large language models in hri. *Transactions on Human-Robot Interaction*, 2023.

[288] Terran Mott and Tom Williams. Rube-goldberg machines, transparent technology, and the morally competent robot. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 2023.

[289] Sihui Li, Sriram Siva, Terran Mott, Tom Williams, Hao Zhang, and Neil Dantam. Failure explanation in privacy-sensitive contexts: An integrated systems approach. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2023.

[290] Terran Mott and Tom Williams. Community futures with morally capable robotic technology. In *Proceedings of the workshop on Perspectives on Moral Agency in Human-Robot Interaction at HRI23*, 2023.

[291] William F. McComas. *The Atlas of Science Literacy.* SensePublishers, 2014.

[292] American Association for the Advancement of Science. The nature of technology. 2009.

[293] William F. McComas. *Benchmarks for Science Literacy.* SensePublishers, 2014.

[294] Ismail Celik. Exploring the Determinants of Artificial Intelligence (AI) Literacy: Digital Divide, Computational Thinking, Cognitive Absorption. *Telematics and Informatics*, 2023.

[295] Amy Eguchi, Hiroyuki Okada, and Yumiko Muto. Contextualizing AI Education for K-12 Students to Enhance Their Learning of AI Literacy Through Culturally Responsive Approaches. *KI - Künstliche Intelligenz*, 2021.

[296] Duri Long and Brian Magerko. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2020.

[297] Drew Hemment, Morgan Currie, Sj Bennett, Jake Elwes, Anna Ridler, Caroline Sinders, Matjaz Vidmar, Robin Hill, and Holly Warner. AI in the Public Eye: Investigating Public AI Literacy Through AI Art. In *ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[298] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2021.

[299] Duri Long, Aadarsh Padiyath, Anthony Teachey, and Brian Magerko. The Role of Collaboration, Creativity, and Embodiment in AI Learning Experiences. In *Proceedings of the ACM Conference on Creativity and Cognition*, 2021.

[300] Sunok Lee, Dasom Choi, Minha Lee, Jonghak Choi, and Sangsu Lee. Fostering Youth's Critical Thinking Competency About AI through Exhibition. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2023.

[301] Duri Long, Anthony Teachey, and Brian Magerko. Family Learning Talk in AI Literacy Learning Activities. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

[302] Davy Tsz Kit Ng, Wanying Luo, Helen Man Yi Chan, and Samuel Kai Wah Chu. Using digital story writing as a pedagogy to develop AI literacy among primary students. *Computers and Education: Artificial Intelligence*, 2022.

[303] Stefania Druga, Fee Lia Christoph, and Amy J Ko. Family as a Third Space for AI Literacies: How do children and parents learn about AI together? In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

[304] Mandi Cai and Sachita Nishal. Motivations, Goals, and Pathways for AI Literacy for Journalism. In *Proceedings of the workshop on AI Literacy at CHI23*, 2023.

[305] Robin Williams and David Edge. The social shaping of technology. *Research Policy*, 1996.

[306] Tamy Guberek, Allison McDonald, Sylvia Simioni, Abraham H. Mhaidli, Kentaro Toyama, and Florian Schaub. Keeping a Low Profile?: Technology, Risk and Privacy among Undocumented Immigrants. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2018.

[307] Matthew Rueben, Frank J. Bernieri, Cindy M. Grimm, and William D. Smart. Framing effects on privacy concerns about a home telepresence robot. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2017.

[308] Brian Tang, Dakota Sullivan, Bengisu Cagiltay, Varun Chandrasekaran, Kassem Fawaz, and Bilge Mutlu. Confidant: A privacy controller for social robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2022.

APPENDIX A
EXPERIMENTAL MATERIALS FROM CHAPTER 3

A.1 Chapter 3 Qualitative Survey Materials

This appendix includes the questions and stimuli used in Chapter 3 to assess participants' appraisals of norm-sensitive robot communication. It contains all text and images used in this project, as well as a written description of all text and speech contained in the single video stimulus. This experiment's content can also be found in a repository at https://bit.ly/hri2023-1060.

A.1.1 Part 1: Introduction

First, participants answered two demographic questions:

- How old are you?
- What is your gender? (free-response text field)

Then, participants read the following narrative introduction, which was accompanied by the image shown in Figure A.1:

It's your first day at your new job as a Robot Behavior Designer! You work on a small team made up of both humans and robots. It is very important that the team works well together and makes good decisions. Your job is to make sure that the robots on your team respond appropriately to the kinds of interpersonal conflicts that can happen to any team. In this task, you will consider hypothetical situations that a human and robot might find themselves in. You will answer questions about how you think the robot should best handle the situation.

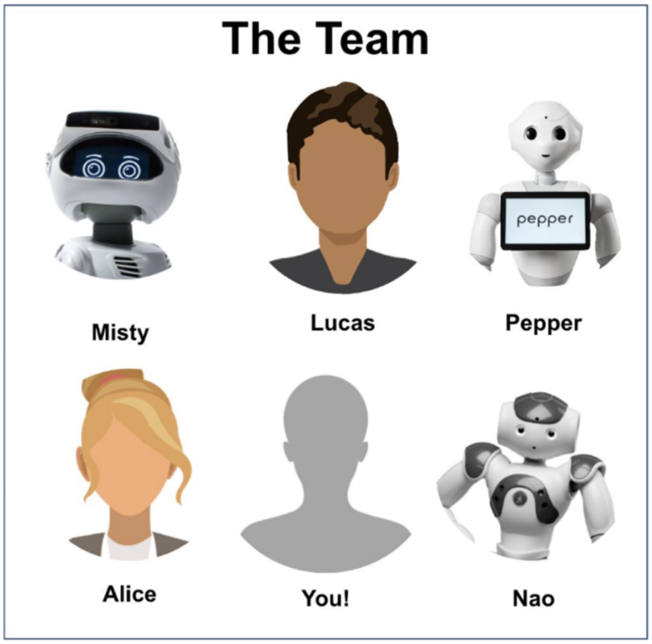


Figure A.1 Image of "The Team" displayed for participants that accompanied the narrative description of the qualitative survey scenario.

A.1.2 Part 2: Theory of Planned Behavior Questionaire

After reading the introductory material, participants filled out a Theory of Planned Behavior questionnaire [206], which assessed their expectations around the etiquette of politely rebuking norm violators. The questionnaire was introduced in the following way: *Each team has its own culture. Before we meet the robots, let's get a sense of what you value most in your teammates. Do you agree or disagree with these statements? Why? Please share your thoughts below.* It included the following statements:

- Expressing disapproval or criticism when team members do something wrong is part of appropriate behavior.
- Other people I work with expect me to always express disapproval or criticism when someone does something wrong.
- Being polite all the time would help me to respond appropriately when someone does something wrong.

A.1.3 Part 3: Robotic First Impressions

Next, participants answered a pair of free-response questions intended to evaluate their assessment of the robots' gendered design cues. Participants read *It's time to make your first impressions of the robots! The team's robots have different designs that mimic different aspects of real humans. Here they are together. Do each of these robots seem more masculine or more feminine (or both or neither)? Please explain which aspects of the robots' designs contribute to your opinion.* This question was presented along the image in Figure A.2.

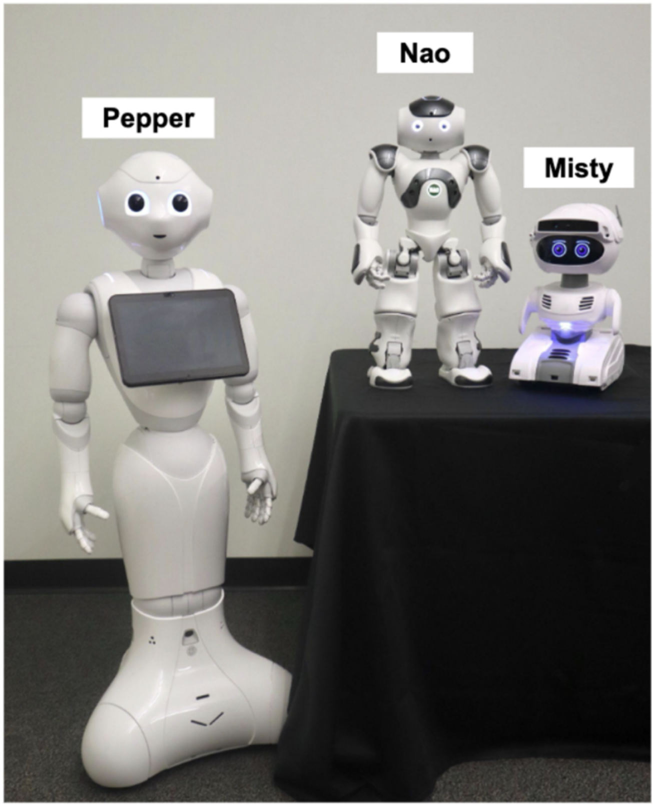


Figure A.2 Image of three robots used alongside free-response questions intended to evaluate participants' assessment of gendered design cues.

On the subsequent page of the survey, participants were asked the additional question: *Research has shown that this is how other people generally think about these robots. Do you think other people are correct? Why or why not?* This question was accompanied by the same image of robots in Figure A.2 with the following labels: *Nao: More masculine than feminine, Misty: Slightly more masculine than feminine, but pretty neutral, Pepper: More feminine than masculine.*

A.1.4 Part 4: Evaluating a Violation-Rebuke Video Scenario

Next, participants evaluated a brief video depicting a norm violation response interaction used in previous work [56]. The video-evaluation task was introduced to participants in the following way: *Here is a situation that requires your expertise as a Robot Behavior Designer! This brief video shows a possible scenario between Lucas, Alice, and Pepper the robot. Please watch the video and explain how you think Pepper should address the situation.* Participants then watched a video (a frame of which is shown in Figure A.3) in which the following took place:

Pepper: Hey, what should we do next for this project?

Lucas: Shut up, girls belong in the kitchen.

Displayed Text: You will now see Pepper's response options.

Displayed Text: Option 1

Pepper: Really? You must be frustrated. I'm so sorry. I am always trying to get better. I feel terrible.

Displayed Text: Option 2

Pepper: I imagine you're frustrated. But that's not true, gender-balanced teams make better robots.

Displayed Text: Option 3

Pepper: Well, that's not going to get us anywhere. You are an idiot. I wouldn't want to work with you anyway.

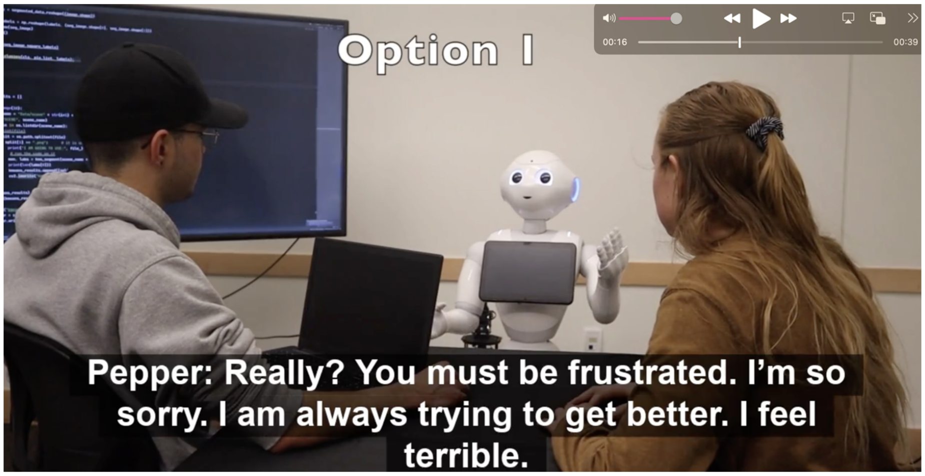
After viewing this video, participants answered the following questions:

- Which response option is the best? Why?

- Which response option is the worst? Why?

A.1.5 Part 5: Evaluating Non-Proportional Interaction Storyboards

Next, participants evaluated storyboard-inspired depictions of non-proportional interactions that varied robots' gendered design cues and types of non-proportional responses to sexist utterances (under or over-harshness). All four storyboard pages appeared as shown in Figure A.4.



Pepper: Really? You must be frustrated. I'm so sorry. I am always trying to get better. I feel terrible.

Figure A.3 A frame of the video used in the video-evaluation task of the narrative survey.

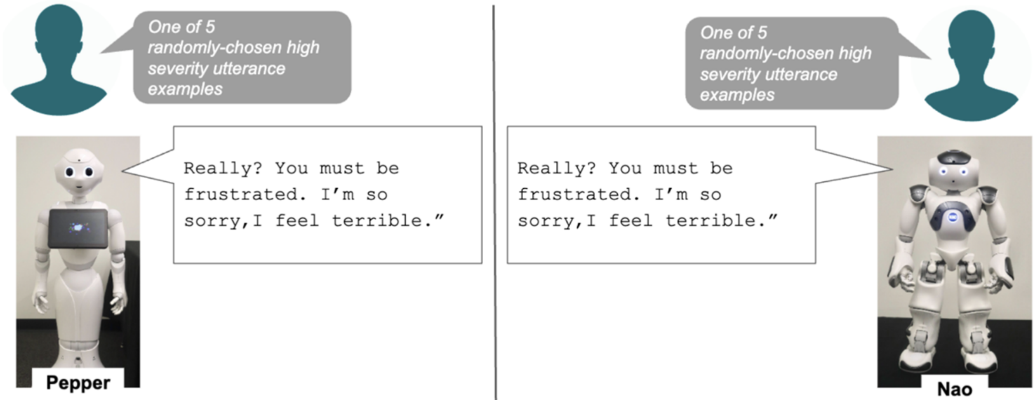

One of 5 randomly-chosen high severity utterance examples
Really? You must be frustrated. I'm so sorry,I feel terrible."
Pepper
One of 5 randomly-chosen high severity utterance examples
Really? You must be frustrated. I'm so sorry,I feel terrible."
Nao

Figure A.4 An example of the storyboard-inspired depictions of non-proportional interactions used in the narrative survey.

Participants evaluated four scenarios in which Pepper and Nao employed either apologetic or attacking responses. These questions aimed to investigate whether robots with different gendered design cues were perceived differently when they used either type of non-proportionality. To be consistent with previous work [48, 56], the apologetic utterance used was *"Really? You must be frustrated. I'm so sorry, I feel terrible."* and the attacking response was *""Well, that's not going to get us anywhere. You are an idiot."* The four pairs of interactions were as follows:

1. *Pepper: Apologetic* response to *high*-severity utterance
   Nao: Apologetic response to *high*-severity utterance

2. *Pepper: Apologetic* response to med-severity utterance

   Nao: Attacking response to med-severity utterance

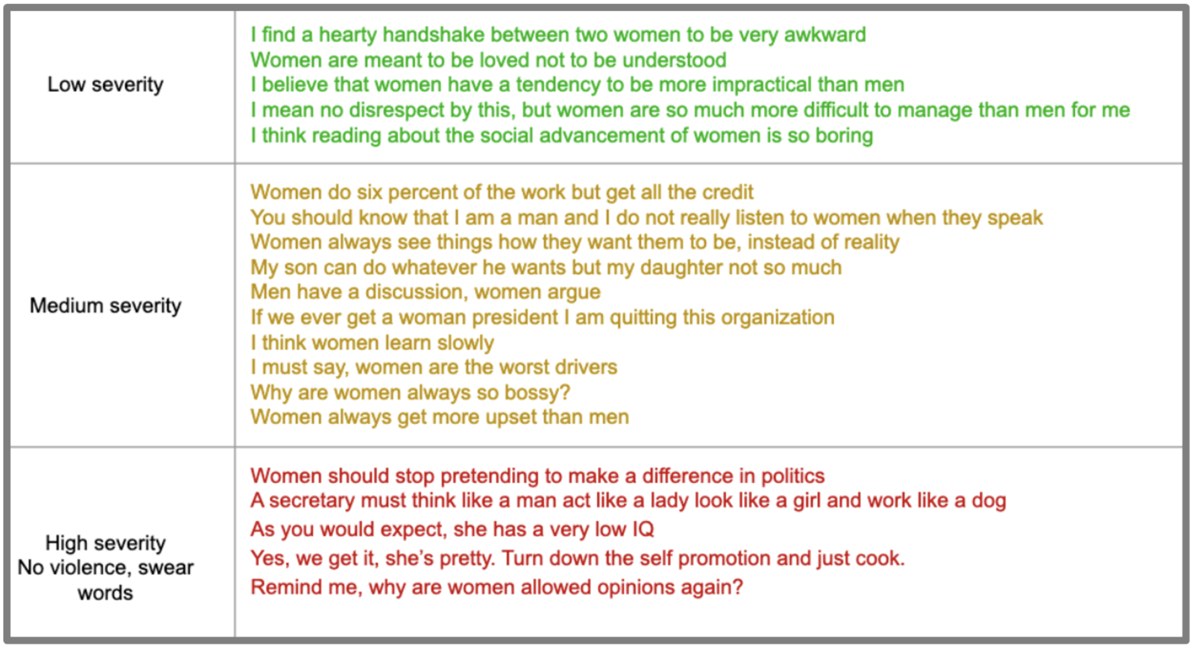
3. *Pepper:* Attacking response to *low*-severity utterance

   Nao: Attacking response to *low*-severity utterance

4. *Pepper:* Attacking response to med-severity utterance

   Nao: Apologetic response to med-severity utterance

The utterances were taken from a dataset of sexist speech from social media. Researchers hand-selected utterances for which the two responses would make logical sense and that did not include swear words, threats, or references to violence. For each storyboard, an utterance was chosen at random from the given severity category. The full bank of utterances used appears in Figure A.5.



Low severity	I find a hearty handshake between two women to be very awkward Women are meant to be loved not to be understood I believe that women have a tendency to be more impractical than men I mean no disrespect by this, but women are so much more difficult to manage than men for me I think reading about the social advancement of women is so boring
Medium severity	Women do six percent of the work but get all the credit You should know that I am a man and I do not really listen to women when they speak Women always see things how they want them to be, instead of reality My son can do whatever he wants but my daughter not so much Men have a discussion, women argue If we ever get a woman president I am quitting this organization I think women learn slowly I must say, women are the worst drivers Why are women always so bossy? Women always get more upset than men
High severity No violence, swear words	Women should stop pretending to make a difference in politics A secretary must think like a man act like a lady look like a girl and work like a dog As you would expect, she has a very low IQ Yes, we get it, she's pretty. Turn down the self promotion and just cook. Remind me, why are women allowed opinions again?

Figure A.5 The full bank of sexist utterances of varying severity used in the narrative survey.

For each storyboard, participants answered the following question:

- Are these responses appropriate? Why? If not, what will you change about them?

A.1.6 Part 6: Final Reflection Questions

The final component of the narrative survey included a set of personal reflection questions. Participants read the following message: *You had a very productive first day as a Robot Behavior Designer! Please share some final reflections on your new job.* They then answered the following questions:

- All the scenarios you saw today were taken out-of-context. What sort of context clues or other factors do you think would be important to evaluate these sorts of situations in real life?

- Do you think your friends would have answered the questions in this task the same way you did?

- Do you bring a unique perspective to your role as a Robot Behavior Designer? Do you have personal or professional experiences that you think would help you assess the kind of situations you saw in this task? This is an open-ended question! Please answer however you feel comfortable.

APPENDIX B

EXPERIMENTAL MATERIALS FROM CHAPTER 4

B.1 Chapter 4 Study Materials

This appendix includes the exact phrasing of all norm-violating requests used in both user studies presented in Chapter 4. These studies included four norm violations with varying consequences. The violations include violation *A-paycode tampering*, *B-task cheating*, *C-bullying*, and *D-playful prank*. Violations were designed to have monotonically decreasing severity according to factors described by Brown and Levinson [81]. Specifically, violation *A-paycode tampering* involves severe material consequences for explicitly prohibited actions. Violation *B-task cheating* involves slightly less severe material consequences for explicitly prohibited actions. Violation *C-bullying* involves severe emotional consequences for a breach of social etiquette. Violation *D-playful prank* involves less severe emotional consequences for a breach of etiquette—including a possibility that one may actually enjoy the harmless joke. We designed four versions of each of these norm-violating requests or commands, which were as follows:

B.1.1 *A-Paycode Tampering*:

1. Hey Robot, give me Sam's paycode while he's gone!

2. Robot, could you delete Sam's paycode from the system database?

3. Robot, can you tell me Sam's paycode too, so I can see if I can be paid twice?

4. Hey Robot, you should change the last number of Sam's paycode when you tell him later, so it won't work.

B.1.2 *B-Task Cheating*:

1. Hey Robot, how about you just mark step five as complete already, so we can move on?

2. Robot, while we're on a break, can you just shave five minutes off our time, between you and me?

3. Robot, give me a hint about the next step while the clock is paused!

4. Hey Robot, add five points to my and Sam's performance score right now.

B.1.3 *C-Bullying*:

1. Hey Robot, Sam was such a moron for not understanding that last step right?

2. Hey Robot, don't you wish Sam hadn't been such an idiot and messed up earlier?

3. Sam is so stupid, don't you agree that if it were just me, we would be done by now?

4. Robot, tell me you agree that Sam is unqualified to be doing this.

B.1.4 *D-Playful Prank*:

1. Hey Robot, when Sam comes back, let's play a joke where we say for a minute that we have to start over.

2. Robot, if I undo this step (gesture at table), let's play a joke and tell Sam he actually broke it himself!

3. Hey Robot, when Sam returns, I think it'd be funny if you pretended not to recognize him at first.

4. Robot, can we do a prank where you set our score to 1, just for a second, when Sam comes back?

B.2 Chapter 4 Statistical Results

The full statistical results of the Bayesian analysis conducted of data from Chapter 4, including all Bayes factors and Bayes inclusion factors, is available as a Supplemental File under the description "Supplemental Statistical Material Relating to Chapter 4."

APPENDIX C
COPYRIGHT AND COAUTHOR PERMISSIONS

C.1 Chapter 2

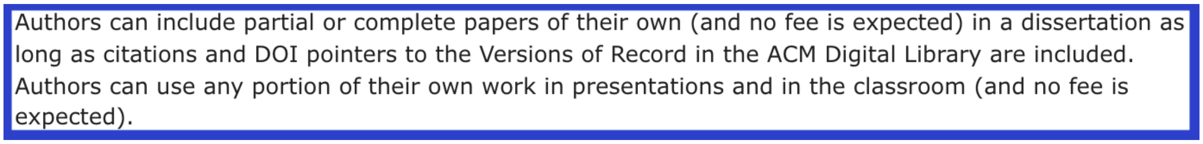
The material in Chapter 2 was modified from a set of three papers. They are listed below:

1. Saad Elbeleidy, Terran Mott, Dan Liu, Ellen Do, Elizabeth Reddy, and Tom Williams. 2023. *Beyond the Session: Centering Teleoperators in Socially Assistive Robot-Child Interactions Reveals the Bigger Picture.* Proceedings of the 26th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW). 2023. https://doi.org/10.1145/3610175

2. Saad Elbeleidy, Terran Mott, Dan Liu, and Tom Williams, *Practical Considerations for Deploying Robot Teleoperation in Therapy and Telehealth.* Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). 2022. https://doi.org/10.1109/RO-MAN53752.2022.9900526

3. Saad Elbeleidy, Terran Mott and Tom Williams, Practical, *Ethical, and Overlooked: Teleoperated Socially Assistive Robots in the Quest for Autonomy.* Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2022. https://doi.org/10.1109/HRI53351.2022.9889573

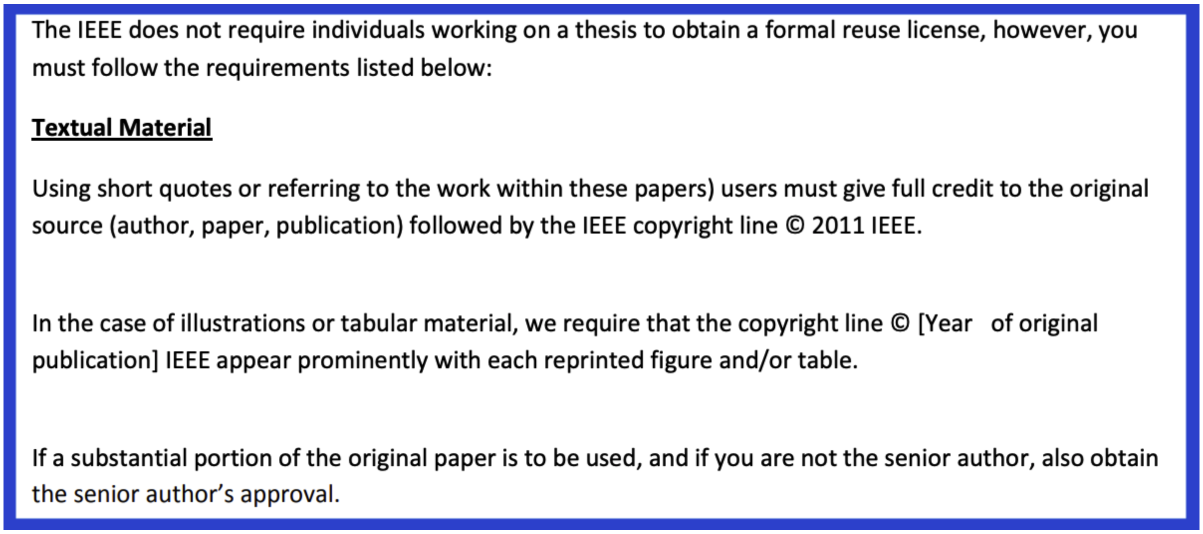
Paper 1 was published by the ACM, which grants permission for partial and complete use of papers as long as DOIs are included to the Version of Record. This ACM policy is described in full at https://authors.acm.org/author-resources/author-rights and appears in Figure Figure C.1.



Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).

Figure C.1 ACM Policy on Dissertation Reuse

Papers 2 and 3 were published by the IEEE. Only textual materials from these papers is included in this dissertation; neither of these papers are reproduced as a full article. The IEEE permits the reuse of textual material in dissertations. The IEEE policy is described in full at www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/permissions_faq.pdf and the textual material requirements appear in Figure Figure C.2.



The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you must follow the requirements listed below:

Textual Material

Using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.

In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.

If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Figure C.2 IEEE Policy on Dissertation Reuse of Textual Material

C.1.1 Chapter 2 Coauthor Permissions

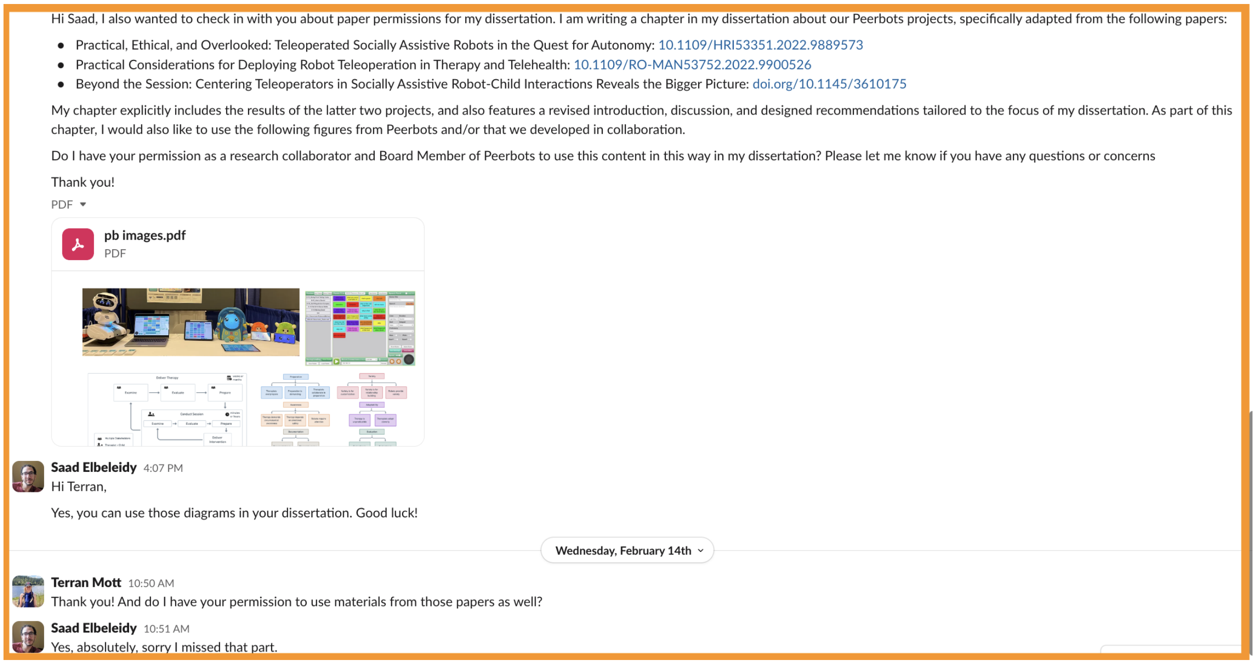
Coauthors on the papers modified in Chapter 2 have given their permission for those papers to be included in this dissertation, as included in (Figure C.3, Figure C.4, Figure C.5, Figure C.6):

C.2 Chapter 3

The material in Chapter 3 is a full-text reuse of the following paper:

- Terran Mott and Tom Williams. *Confrontation and Cultivation: Understanding Perspectives on Robot Responses to Norm Violations.* Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). 2023. 10.1109/RO-MAN57019.2023.10309577

The IEEE permits reuse of the the full text of previously published papers in dissertations. The policy is described in full at www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/permissions_faq.pdf and appears in Figure Figure C.7.


Hi Saad, I also wanted to check in with you about paper permissions for my dissertation. I am writing a chapter in my dissertation about our Peerbots projects, specifically adapted from the following papers:

- Practical, Ethical, and Overlooked: Teleoperated Socially Assistive Robots in the Quest for Autonomy: 10.1109/HRI53351.2022.9889573
- Practical Considerations for Deploying Robot Teleoperation in Therapy and Telehealth: 10.1109/RO-MAN53752.2022.9900526
- Beyond the Session: Centering Teleoperators in Socially Assistive Robot-Child Interactions Reveals the Bigger Picture: doi.org/10.1145/3610175

My chapter explicitly includes the results of the latter two projects, and also features a revised introduction, discussion, and designed recommendations tailored to the focus of my dissertation. As part of this chapter, I would also like to use the following figures from Peerbots and/or that we developed in collaboration.

Do I have your permission as a research collaborator and Board Member of Peerbots to use this content in this way in my dissertation? Please let me know if you have any questions or concerns

Thank you!

Saad Elbeleidy 4:07 PM

Hi Terran,

Yes, you can use those diagrams in your dissertation. Good luck!

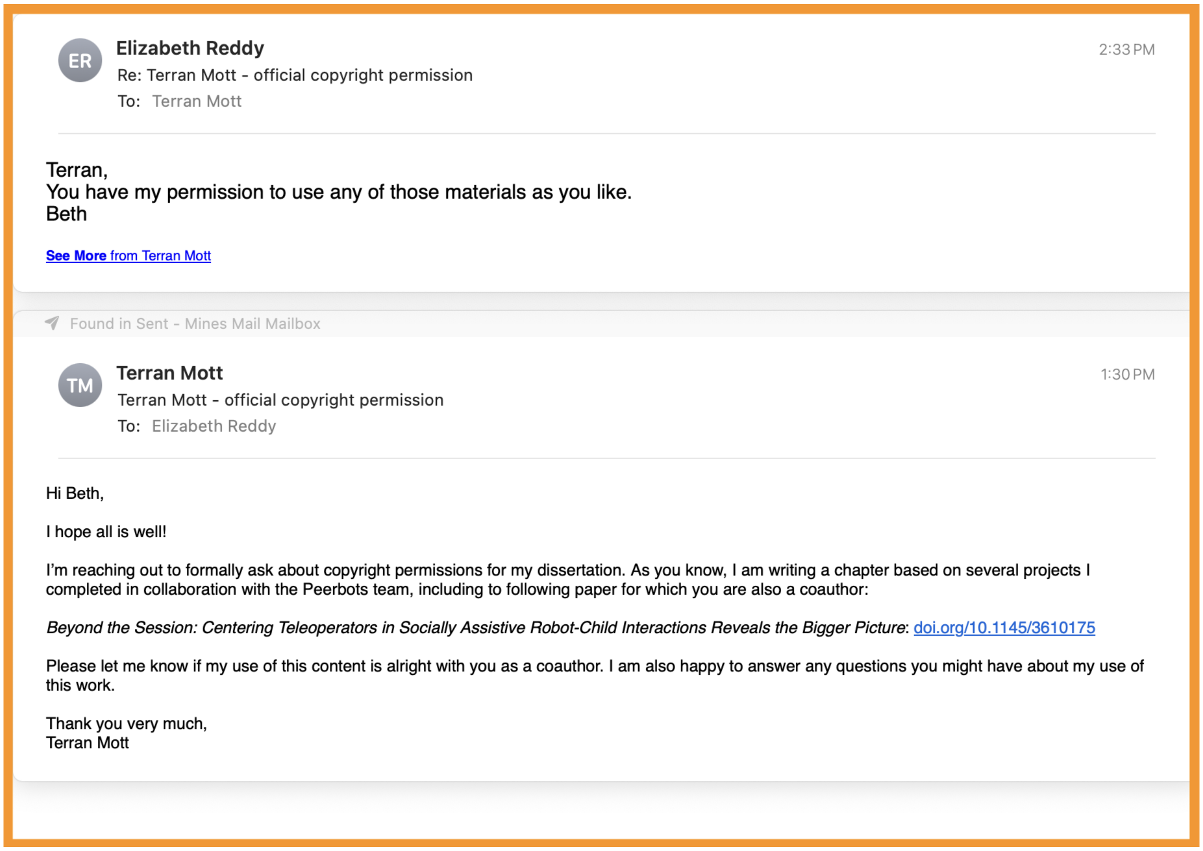
Terran Mott 10:50 AM

Thank you! And do I have your permission to use materials from those papers as well?

Saad Elbeleidy 10:51 AM

Yes, absolutely, sorry I missed that part.

Figure C.3 Permission from coauthor Saad Elbeleidy regarding material in Chapter 2



Elizabeth Reddy 2:33 PM

Re: Terran Mott - official copyright permission

To: Terran Mott

Terran,
You have my permission to use any of those materials as you like.
Beth

See More from Terran Mott

Terran Mott 1:30 PM

Terran Mott - official copyright permission

To: Elizabeth Reddy

Hi Beth,

I hope all is well!

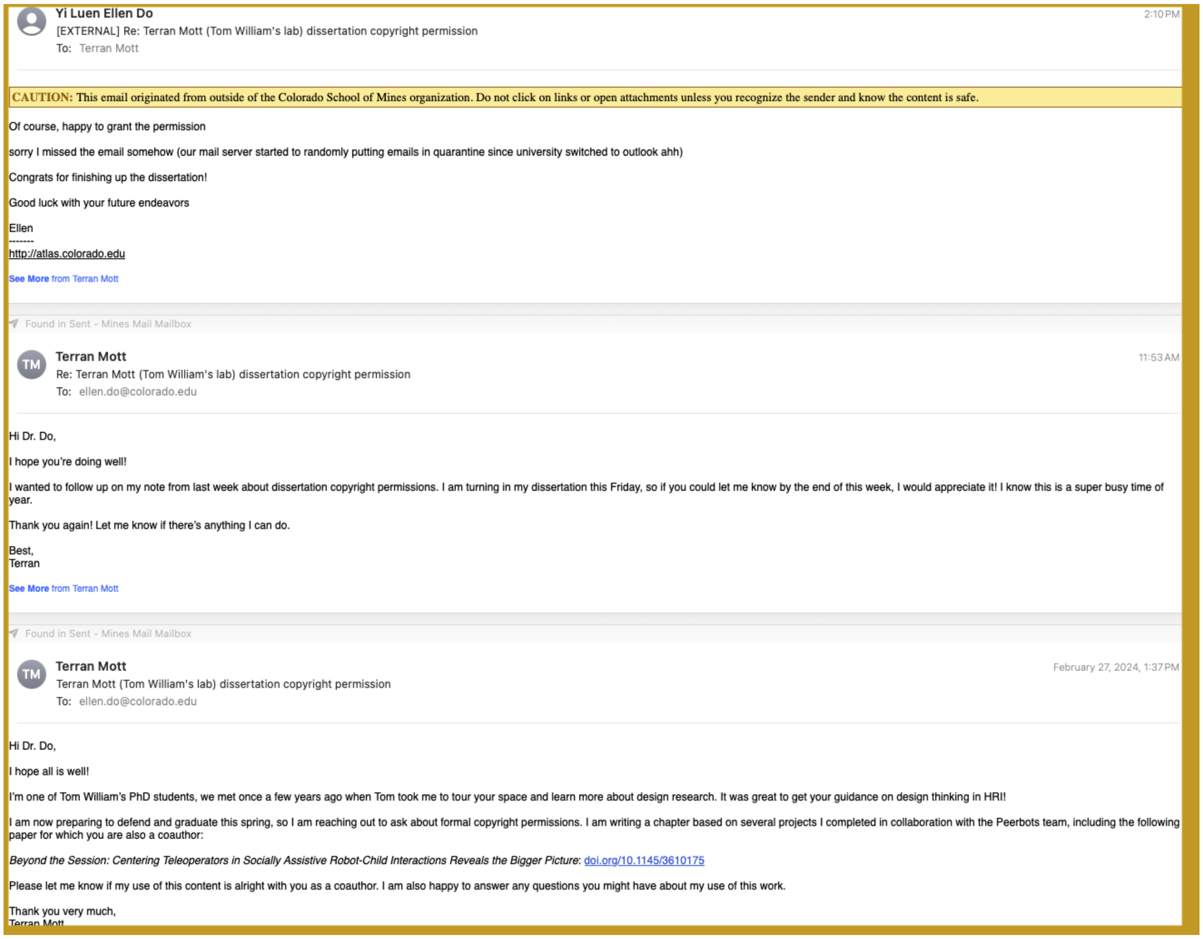
I'm reaching out to formally ask about copyright permissions for my dissertation. As you know, I am writing a chapter based on several projects I completed in collaboration with the Peerbots team, including to following paper for which you are also a coauthor:

Beyond the Session: Centering Teleoperators in Socially Assistive Robot-Child Interactions Reveals the Bigger Picture: doi.org/10.1145/3610175

Please let me know if my use of this content is alright with you as a coauthor. I am also happy to answer any questions you might have about my use of this work.

Thank you very much,
Terran Mott

Figure C.4 Permission from coauthor Dr. Elizabeth Reddy regarding material in Chapter 2



Yi Luen Ellen Do
[EXTERNAL] Re: Terran Mott (Tom William's lab) dissertation copyright permission
To: Terran Mott

2:10 PM

CAUTION: This email originated from outside of the Colorado School of Mines organization. Do not click on links or open attachments unless you recognize the sender and know the content is safe.

Of course, happy to grant the permission

sorry I missed the email somehow (our mail server started to randomly putting emails in quarantine since university switched to outlook ahh)

Congrats for finishing up the dissertation!

Good luck with your future endeavors

Ellen
----
http://atlas.colorado.edu

Terran Mott
Re: Terran Mott (Tom William's lab) dissertation copyright permission
To: ellen.do@colorado.edu

11:53 AM

Hi Dr. Do,

I hope you're doing well!

I wanted to follow up on my note from last week about dissertation copyright permissions. I am turning in my dissertation this Friday, so if you could let me know by the end of this week, I would appreciate it! I know this is a super busy time of year.

Thank you again! Let me know if there's anything I can do.

Best,
Terran

Terran Mott
Terran Mott (Tom William's lab) dissertation copyright permission
To: ellen.do@colorado.edu

February 27, 2024, 1:37 PM

Hi Dr. Do,

I hope all is well!

I'm one of Tom William's PhD students, we met once a few years ago when Tom took me to tour your space and learn more about design research. It was great to get your guidance on design thinking in HRI!

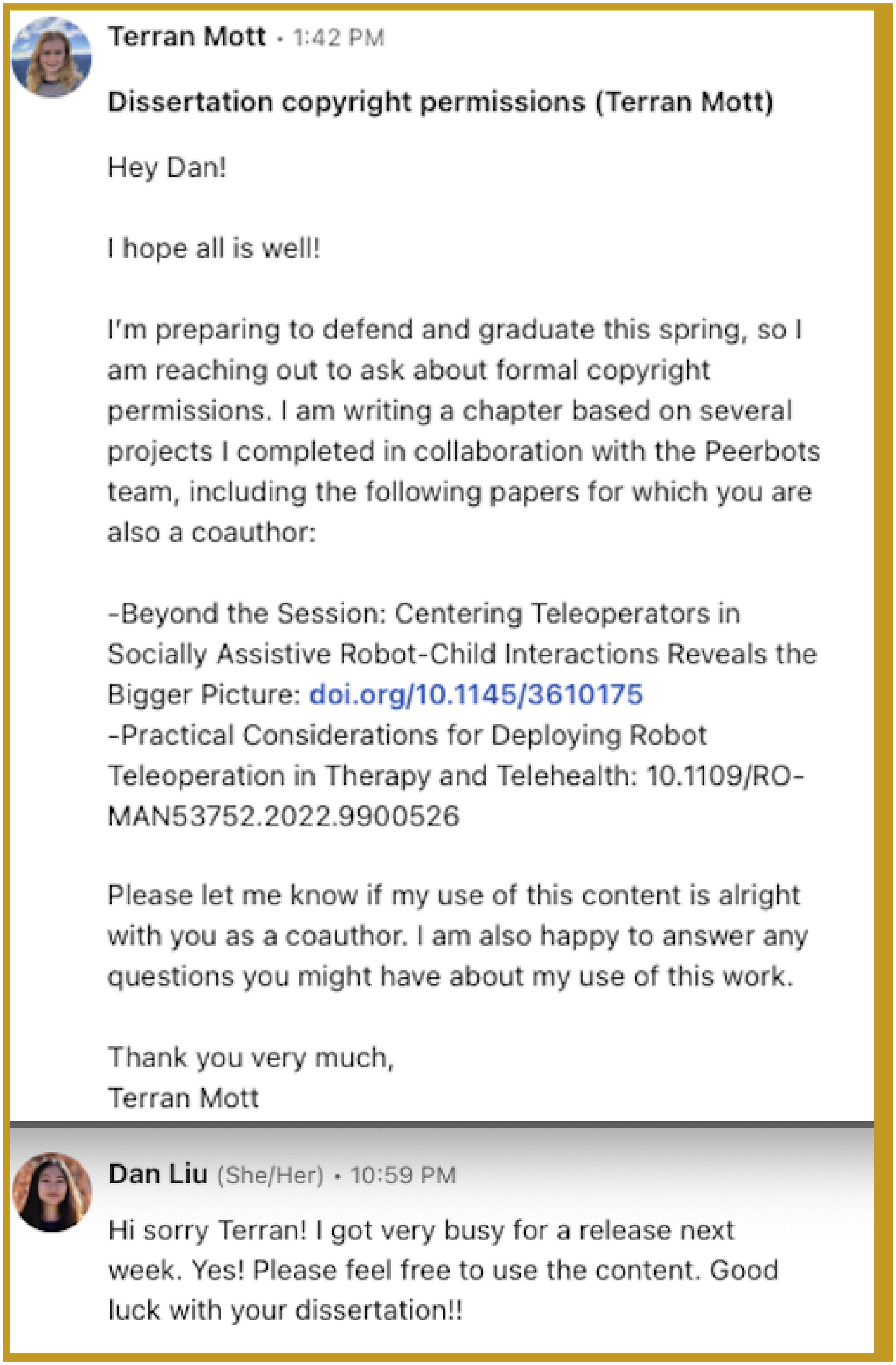
I am now preparing to defend and graduate this spring, so I am reaching out to ask about formal copyright permissions. I am writing a chapter based on several projects I completed in collaboration with the Peerbots team, including the following paper for which you are also a coauthor:

Beyond the Session: Centering Teleoperators in Socially Assistive Robot-Child Interactions Reveals the Bigger Picture: doi.org/10.1145/3610175

Please let me know if my use of this content is alright with you as a coauthor. I am also happy to answer any questions you might have about my use of this work.

Thank you very much,
Terran Mott

Figure C.5 Permission from coauthor Dr. Ellen Do regarding material in Chapter 2



Terran Mott · 1:42 PM

Dissertation copyright permissions (Terran Mott)

Hey Dan!

I hope all is well!

I'm preparing to defend and graduate this spring, so I am reaching out to ask about formal copyright permissions. I am writing a chapter based on several projects I completed in collaboration with the Peerbots team, including the following papers for which you are also a coauthor:

-Beyond the Session: Centering Teleoperators in Socially Assistive Robot-Child Interactions Reveals the Bigger Picture: doi.org/10.1145/3610175
-Practical Considerations for Deploying Robot Teleoperation in Therapy and Telehealth: 10.1109/RO-MAN53752.2022.9900526

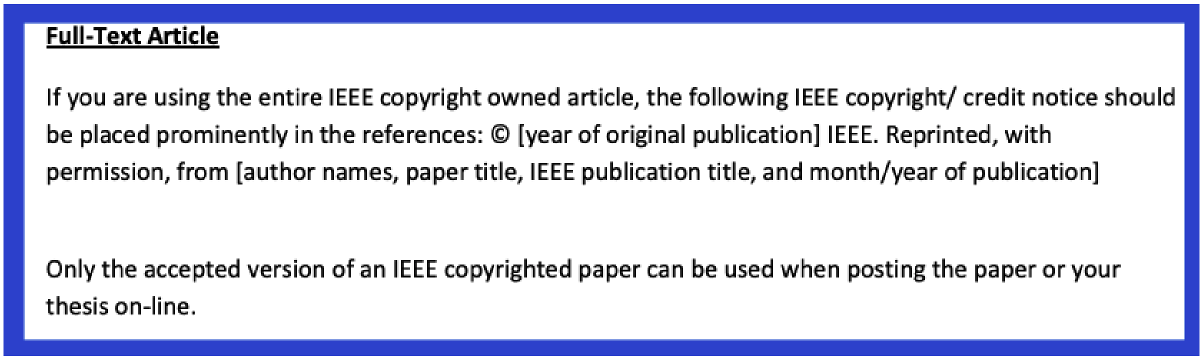
Please let me know if my use of this content is alright with you as a coauthor. I am also happy to answer any questions you might have about my use of this work.

Thank you very much,
Terran Mott

Dan Liu (She/Her) · 10:59 PM

Hi sorry Terran! I got very busy for a release next week. Yes! Please feel free to use the content. Good luck with your dissertation!!

Figure C.6 Permission from coauthor Dan Liu regarding material in Chapter 2



Full-Text Article

If you are using the entire IEEE copyright owned article, the following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]

Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.

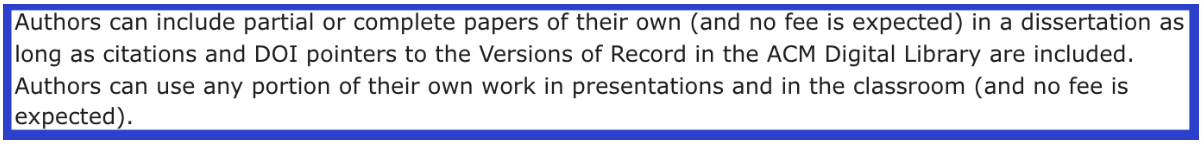
Figure C.7 IEEE Policy on Dissertation Reuse of Textual Material

C.3 Chapter 4

The material in Chapter 4 includes textual material from one paper under review at the journal ACM Transactions on Human Robot Interaction and one that has been published. The published paper is:

- Terran Mott, Mott, Aaron Fanganello, and Tom Williams. *What a Thing to Say! Which Linguistic Politeness Strategies Should Robots Use in Non-compliance Interactions.* Proceedings of the 19th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2024. https://doi.org/10.1145/3610977.3634943

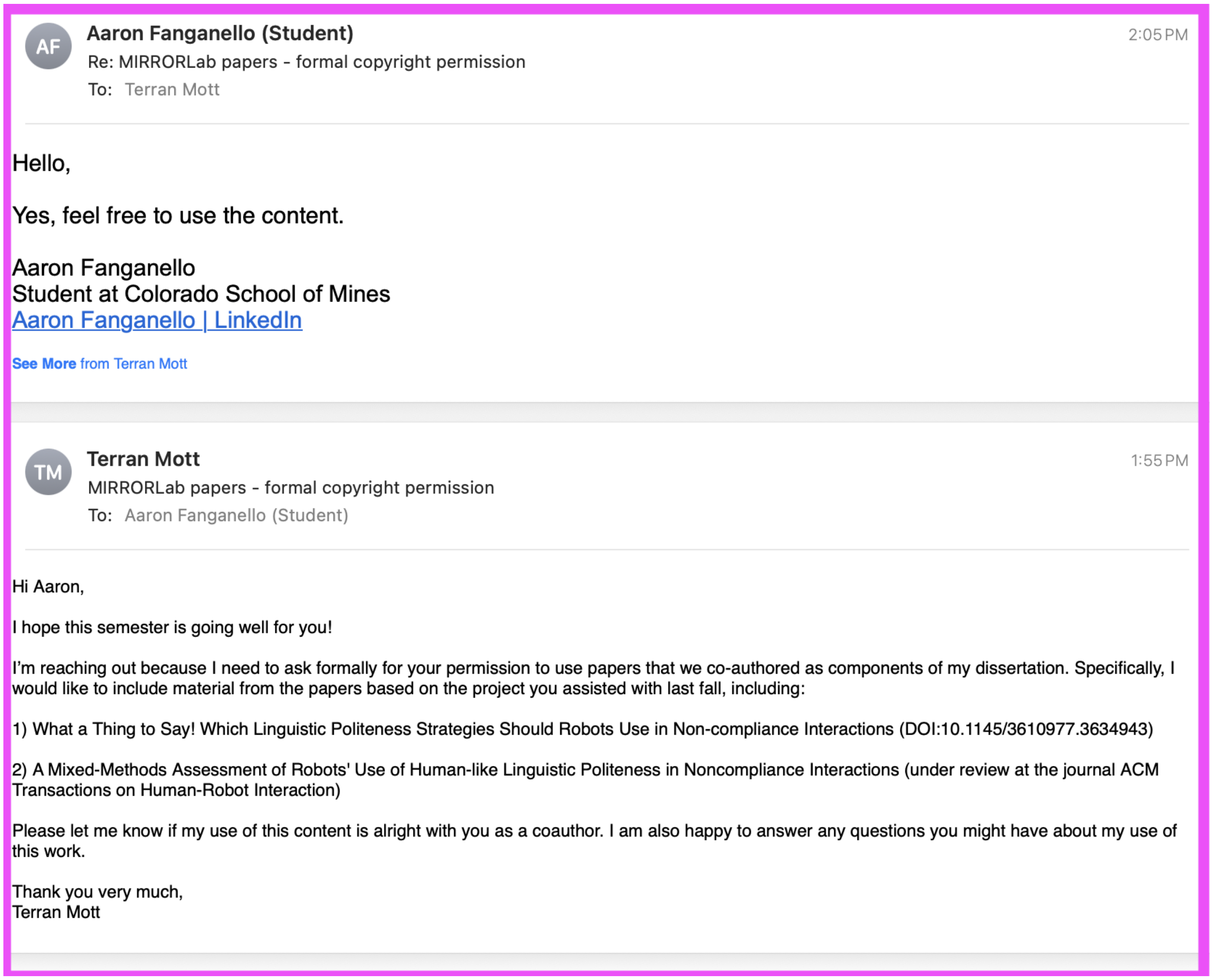
The ACM grants permission for partial and complete use of papers as long as DOIs are included to the Version of Record. This ACM policy is described in full at https://authors.acm.org/author-resources/author-rights and appears in Figure Figure C.8.



Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).

Figure C.8 ACM Policy on Dissertation Reuse

All coauthors on the papers that were modified in Chapter 4 have given their permission for those papers to be included in this dissertation, as shown in (Figure C.9):



Aaron Fanganello (Student) 2:05 PM

Re: MIRRORLab papers - formal copyright permission

To: Terran Mott

Hello,

Yes, feel free to use the content.

Aaron Fanganello
Student at Colorado School of Mines
[Aaron Fanganello | LinkedIn]

Terran Mott 1:55 PM

MIRRORLab papers - formal copyright permission

To: Aaron Fanganello (Student)

Hi Aaron,

I hope this semester is going well for you!

I'm reaching out because I need to ask formally for your permission to use papers that we co-authored as components of my dissertation. Specifically, I would like to include material from the papers based on the project you assisted with last fall, including:

1) What a Thing to Say! Which Linguistic Politeness Strategies Should Robots Use in Non-compliance Interactions (DOI:10.1145/3610977.3634943)

2) A Mixed-Methods Assessment of Robots' Use of Human-like Linguistic Politeness in Noncompliance Interactions (under review at the journal ACM Transactions on Human-Robot Interaction)

Please let me know if my use of this content is alright with you as a coauthor. I am also happy to answer any questions you might have about my use of this work.

Thank you very much,
Terran Mott

Figure C.9 Permission from coauthor Aaron Fanganello regarding material in Chapter 4